Retrieval Monitoring Affects Self-Regulated Learning of Young and Older Adults

A thesis submitted by

Renée DeCaro

in partial fulfillment of the requirements for the degree of

Master of Science

in

Psychology

Tufts University

May 2018

Advisor: Ayanna K. Thomas

Abstract

Successful learning and remembering in older adulthood is essential, and its failures can have uniquely serious consequences. The present research investigates failures that arise from metamemorial deficits, and tests their influence on how one regulates subsequent learning. The present study tested the influence of retrieval and cues derived from the structure of the task on self-regulated learning in older and young adults. In three experiments, young and older adults studied unrelated cue-target pairs, and made item-by-item monitoring judgments after initial study had concluded. In Experiments 1 and 2, participants made feeling of knowing (FOK) predictions, after attempted target retrieval. Participants engaged in control, implemented as the decision to restudy a subset of items, either following the initial monitoring phase (Experiments 1 and 3) or during the initial monitoring phase (Experiment 2). Participants restudied the selected cue-target pairs and received a final recall test on all items. In Experiment 1, both age groups demonstrated a negative relationship between FOK and control, though the relationship was stronger in young compared to older adults. Experiments 2 and 3 established how specific metamemorial task designs promoted or deterred age differences. When control was implemented during the monitoring phase on a trial-by-trial basis (Experiment 2), the relationship between monitoring and control was weakened. When monitoring instructed a specific retrieval attempt (Experiment 3), absolute confidence was lower than when individuals were not instructed to attempt retrieval. These results suggest that retrieval and monitoring predict whether an item will be restudied later in both younger and older adults, and that observed age differences may depend, in part, on design features within the experimenter's control.

TABLE OF CONTENTS

Introduction **1**
Retrieval as the Basis for Monitoring 4
Metamemorial Task Designs and Self-Regulated Learning 8
The Present Study 11
Experiment 1 ..... **13**
Method 13
Participants 13
Materials 14
Procedure 14
Results 17
Factors Influencing Monitoring and Control 17
Selection of Items for Restudy 21
Mixed-Effects Model 22
Discussion 26
Experiment 2 ..... **28**
Method 29
Participants 29
Procedure 29
Results 30
Factors Influencing Monitoring and Control 30
Mixed-Effects Model 32
Discussion 34
Experiment 3 ..... **34**
Method 36
Design 36
Participants 37
Materials 37
Procedure 38
Results 40
Factors Influencing Monitoring and Control 40
Selection of Items for Restudy 46
Final Test 49
Discussion 50
General Discussion **52**
Retrieval as the Basis for Monitoring 53
Impact of the Demands of the Task 57
Conclusions 58

LIST OF FIGURES

Experiment 1 .....**13**
Figure 1: Average feeling of knowing rating by item selection.....21
Figure 2: Predicted probability of selecting at item to be restudied.....25
Experiment 2 .....**28**
Figure 3: Predicted probability of selecting at item to be restudied.....33
Experiment 3 .....**34**
Figure 4: Predicted reaction times (in seconds) across units of Level 1 JOL 44
Figure 5: Mean reaction times (RT) in each of the JOL categories 64
Figure 6: Average judgment of learning rating by item selection.....47
Figure 7: Frequency distribution of the use of each of the JOL categories 48

Retrieval Monitoring Affects Self-Regulated Learning of Young and Older Adults

Successful learning and remembering in older adulthood facilitates independent living, allowing older adults to adapt to lifestyle changes and increasingly complicated health care regimens. As one example, an older adult may need to remember to take several medications on different schedules—for instance she must take her medication *Toprol* with food and her medication *Restoril* at night. She may monitor how well she remembers the information about her medications and, depending on how well she thinks she knows what she needs to remember, she may adopt a strategy for learning (such as creating a mnemonic i.e. *Restoril* contains the word 'rest' so take at night).

For this older adult, failures in learning and remembering can have serious consequences. She may forget to take a needed medication or take a medication twice—either of which could cause dire effects to her health. Problems may be derived when she does not accurately gauge her memory for the information; or, she may know what she knows and what she does not know, but may not implement an appropriate strategy to effectively learn unknown information. Importantly, these errors can originate from multiple sources.

From a cognitive aging perspective, age-related deficits may be attributable to declines in available attentional or executive function resources. Declines in these resources (e.g. processing speed) lead to observed age-differences in specific memory functions such as recalling recently learned information, recognizing incidentally encoded information, or learning new associations (see Craik, 2000). From a compensatory perspective, a standard metacognitive framework (e.g. Nelson & Narens, 1990) may be useful in mitigating declines in these memory functions (Hertzog, 2016).

As such, the focus of the present research is to investigate failures that arise from metamemorial deficits.

Metamemory is thought to be comprised of two interacting components: monitoring and control (e.g. Nelson & Narens, 1990). Monitoring involves the awareness and assessment of current states of learning and retrieval, and may also involve predictions regarding learning and retrieval. Control involves the regulation of learning, and includes decisions about what information should be learned and when and how to learn that information (for review, see Dunlosky, Serra, & Baker, 2007).

Most research examining the relationship between monitoring and control proposes that monitoring affects control (but see Koriat, Ma'ayan, & Nussinson, 2006). That is, learners use the outcome of monitoring processes to make decisions (*control*), such as restudying or stopping study of the to-be-learned information. Thus, *self-regulated learning* involves monitoring one's learning (i.e. determining whether an item has been adequately learned relative to some learning goal) and the behavioral consequences of that monitoring, such as deciding which items to study and in what order. To this effect, models of self-regulated learning have been advanced to account for the decisions one makes when regulating learning.

One model predicts that learners will focus first and longest on unknown, or the least well known, items rather than known, or better known, items. The purpose of this strategy is to reduce the discrepancy or difference between the learner's *desired* level of knowledge and her *current* state of knowledge. Her selection behavior is in line with this *discrepancy reduction model* (Dunlosky & Hertzog, 1998; Nelson & Leonesio, 1988) when a negative relationship is shown between control (i.e. study time allocation) and

monitoring (i.e. confidence judgments). There is much empirical support for this model. In a comprehensive review, Son and Metcalfe (2000) reported that out of 46 published experimental conditions, 35 supported the discrepancy reduction model's predictions.

However, learners do not always focus on the least well known items. When study time is limited (Metcalfe, 2002) or learning goals are low (Theide & Dunlosky, 1999) learners may first select easier unknown items. In these cases, the effectiveness of decisions during study will depend on the individual: for a given individual, the optimal strategy is to study the subset of items closest to one's level of current mastery. For that reason, this model is called *region of proximal learning* (e.g. Kornell & Metcalfe, 2006; Metcalfe, 2002). Finally, in instances where the value of the to-be-learned items is emphasized, learners may choose to focus on items in accordance with the item's respective worth or importance (e.g. Castel, Benjamin, Craik, & Watkins, 2002) in line with a model called *agenda-based regulation*.

Though these models have primarily been tested with younger adults and children (see though Dunlosky & Hertzog, 1997), interest in examining how self-regulated learning changes in older adulthood is increasing. Research studies focusing on older adults' ability to control learning have primarily been relegated to the self-pacing domain, that is, how long people persist in studying a given item, and have examined item-by-item monitoring judgments. The accuracy of these judgments is typically determined using correlations between metacognitive judgments and an outcome measure (e.g. study time allocation, final test performance).

These studies have consistently indicated that older adult learners do not allocate study time in accordance with monitoring judgments to the same extent as young adults

(Dunlosky & Connor, 1997; Miles & Stine-Morrow, 2004; Souchay & Isingrini, 2004; Price & Murray, 2012, but see Hines, Touron, & Hertzog, 2009). The goal of this work was to determine why monitoring judgments are less likely to affect control decisions in older as compared to younger adults. We suggest that this age-related disconnect between monitoring and control may, in part, be a result of specific metamemorial task designs.

Monitoring judgments and their accuracy are assumed to be based, at least some of the time, on an attempt to retrieve (c.f., Son &Metcalfe, 2005). However, the majority of research investigating age differences in self-regulated learning has not formally assessed retrieval. Objective retrieval, however, may be an important determinant of control. When monitoring is assessed in the context of an explicit retrieval attempt, contextual information retrieved concurrently may have downstream consequences on self-regulated learning. The present study examined how the relationship between retrieval-dependent monitoring and control was influenced by 1) explicit retrieval of contextual information, and 2) the demands of the metacognitive task. Both aims served the purpose of delineating when and where metamemorial deficits could arise in older adults' self-regulated learning.

Retrieval as the Basis for Monitoring

Older adults rely more on, and place greater confidence in, the metacognitive experiences of familiarity and fluency compared to younger adults (Jacoby & Rhodes, 2006); this overreliance and overconfidence may bias subsequent control. For example, reliance on familiarity to guide subsequent control decisions may be misplaced in certain contexts, such as when familiarity is artificially inflated via priming (Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014).

One way to reduce reliance on familiarity at least is to require learners to covertly or overtly attempt retrieval prior to engaging in monitoring (Son & Metcalfe, 2005). For older adults, attempted retrieval may be an important predictor of subsequent control (e.g. Dunlosky & Connor, 1997). Specifically, failures to retrieve may reduce overconfidence (Miller & Geraci, 2014), and, this difference in monitoring magnitude may in turn affect subsequent study behavior. Because the act of attempting retrieval itself may affect the level of confidence in future recallability of information, measuring memory performance at the time of monitoring may be crucial in understanding how retrieval, monitoring, and control interact.

Of the handful of studies examining monitoring and control decisions following an explicit retrieval attempt (Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013; Hanczakowski, et al., 2014; Hines, Touron, & Hertzog, 2009; Littrell, 2011; Souchay & Isingrini, 2004), two examine age differences. These two studies highlight how measuring memory performance at the time of monitoring may be critical to understanding when and where potential age differences in control may arise. Generally, their conclusions support the finding that memory performance is significantly related to subsequent study time allocation for both young and older adults, as more time is allocated to studying items not initially recognized or recalled (Hines, Touron, & Hertzog, 2009; Souchay & Isingrini, 2004).

Souchay and Isingrini (2004) reported average within-person correlations of control (the allocation of study time) with two predictors: retrieval (initially recalling an item) and feeling of knowing (FOK). Older adults did not demonstrate as strong of a relationship between either predictor and subsequent study time allocation as young

adults. Specifically, in both cases younger adults demonstrated a stronger negative relationship (successfully recalling the target and higher relative values of FOK were both associated with less time spent studying). While older adults demonstrated this same pattern (a negative relationship), the relationship was not as pronounced. Consequently, the authors concluded that older adults exhibited a control deficit.

Hines, Touron, and Hertzog (2009) demonstrated that the relationship between recognition confidence judgments and subsequent study time was superior for older adults. Participants underwent two study-test phases, where the first study session consisted of experimenter-paced study of 60 word pairs. A yes/no recognition test followed, during which participants made recognition confidence judgments. The second study session was self-paced (the implementation of control). Older adults exhibited study behavior in line with monitoring (confidence), as they allocated more time to items rated with lower confidence compared to items with higher confidence. These results suggest no control deficit among older adults. The two studies, taken together, highlight how an explicit retrieval attempt may have consequences for subsequent control, and how objectively measuring performance may have consequences for the conclusions regarding age differences.

In the present studies, we chose to focus first on feeling of knowing as its predictions are made in the context of an explicit and experimentally measured retrieval attempt. Additionally, research has consistently demonstrated that FOKs may be based on information associated with the unrecalled target (e.g., Brewer, Marsh, Clark-Foos, & Meeks, 2010; Cook, Marsh, & Hicks, 2006; Thomas, Bulevich, & Dubois, 2011). When an individual cannot successfully retrieve a target, she may still be able to access partial

or contextual information about it. Partial target information is that which is accessed and directly relates to the to-be-remembered target (such as its first letter), whereas contextual information is that which is accessed and not specifically related (such as font color). While the distinction between these types of accessed information may not always be clear (see Schwartz, Pillot, & Bacon, 2014), broadly, FOK allows for the direct examination of access to information about the unrecalled target.

The quality, or accuracy, of retrieved partial information influences the magnitude and prediction accuracy of episodic FOK (Brewer et al., 2010; Hertzog, Dunlosky, & Sinclair, 2010; Hertzog, Fulton, Sinclair, & Dunlosky, 2014; Thomas, Bulevich, & Dubois, 2011, 2012). Thomas and colleagues (2011) examined how the quality of partial information retrieved about initially unrecalled targets impacted FOK judgments in young and older adults, and found that when older adults were first questioned about information about the target word, prior to making FOK judgments, their FOK prediction accuracy improved. By adapting the method used by Thomas et al., we evaluated whether target accessibility itself (accessing partial information) was an important source of feedback for self-regulatory control. Thus, including a retrieval attempt is interesting not only for its ability to enable us to disentangle where age differences may exist in self-regulated learning, but also it is of theoretical interest as it enables us to evaluate how processes that underlie retrieval monitoring (accessing partial information) may be relevant to subsequent control.

In this study series, we were also interested in investigating whether age-related differences in self-regulated learning may, in part, be attributable to specific metamemorial task designs. The first two studies emphasized control and how

its implementation may affect cognitive demands. The last study investigated increasing demands at the time of monitoring as a potential source for age differences.

Metamemorial Task Designs and Self-Regulated Learning

Measuring control as the allocation of study time may lead to issues with interpretability. In particular, study time allocation may indicate a host of processes and be influenced by a number of factors (e.g. Koriat et al., 2006). A variety of factors known to influence study time, such as item difficulty (Koriat et al., 2006, p. 40) or "study effort" (Koriat et al., 2006, p. 44), may predict learning behavior better than study time, per se. More specifically, Koriat et al. (2006) argues that study time allocation may be a data-driven process, where learners base monitoring, specifically the predictions of future recallability, on the effort taken to learn the item. Learners spend time according to what they determine an item "calls for" (p. 41).

In contrast, certain situations, such as differential incentives (Castel, Benjamin, Craik, & Watkins, 2002; Dunlosky & Thiede, 1998) or experimenter-placed limitations on time (Metcalfe, 2002), promote the consideration of goals, that is, how one would aim to complete the task as a whole. Goal-driven processes are likely to be promoted by a superordinate, higher-order stage, in which learners must decide whether to focus on easier or more difficult items within the context of completing the demands of the given task (Thiede & Dunlosky, 1999). It is in this scenario, where goal-driven processes are emphasized, that there is less ambiguity concerning the interpretability of how monitoring affects control. Thus, it might be advantageous to consider a higher-order strategy selection stage, which would include choosing which *subset* of items to study over others (Thiede & Dunlosky, 1999).

A few studies have examined whether age-related differences exist in the choice to restudy a subset of items. Dunlosky and Hertzog (1997) examined restudy choices in young and older adults during multi-trial learning. Young and older participants studied 36 unrelated concrete noun pairs at a fixed rate. Next, participants gave delayed *judgments-of-learning* (JOLs). For each item, the *cue* (the first word of the pair) was presented and participants were asked to estimate confidence in future recallability of the second word when prompted with the first. After each cue-only JOL, participants made restudy decisions. Importantly, participants were limited to selecting half of the items to restudy. Regardless of age, average within-person correlations of JOLs and restudy selection were negative. Both young and older adults selected items to restudy that they had judged to be most poorly learned. From their analyses, the authors suggested that both young and older adults used a "functionally identical algorithm to select items for restudy" (p. 182). More recently, a similar effect was demonstrated when young and older adults instead studied both concrete-concrete and abstract-abstract noun pairs. Participants chose a subset of items to restudy (from an array presenting all cues). Both young and older adults chose to restudy items rated with lower monitoring ratings, abstract-abstract pairs—the objectively more difficult items (Tullis & Benjamin, 2012). Thus, prior research has indicated that when control is operationalized as selecting a subset of items, young and older adults exbibit identical selection behaviors (Dunlosky & Hertzog, 1997; Tullis & Benjamin, 2012).

In Experiment 1, to promote goal-orientation, participants made their selections all at once *after* the initial testing and monitoring phase was completed. During the selection phase, participants could see all possible choices in an array that presented all

studied cues. In Experiment 2, participants made their selections *during* the testing phase. That is, participants were given the option to select whether to restudy a given item on a trial-by-trial basis at the time of monitoring. Participants therefore had to concurrently manage several ongoing processes related to both monitoring (such as whether they successfully retrieved the target or information about the target) and control (such as keeping track of what had already been selected and how many items were left to be selected). Our rationale was this: the process of self-regulatory control is cognitively demanding, and when performed in the context of concurrent demands on working memory, may fall beyond the ability of older adults who demonstrate a decline in available cognitive resources (Salthouse, 2010). We manipulated the cognitive demands of the control task across the first two experiments to examine whether doing so differentially affected older adults.

In the third and final study, we considered how manipulating the demands of the monitoring phase would affect subsequent control. As previously mentioned, requiring learners to covertly or overtly attempt retrieval prior to engaging in monitoring likely reduces reliance on familiarity and fluency (Son & Metcalfe, 2005). Support for this finding is derived from comparisons between conditions where participants are instructed to explicitly attempt retrieval prior to monitoring versus not. The different instructions result in a different pattern of reaction times across values of monitoring. Further, practicing explicit retrieval itself may affect confidence (Miller & Geraci, 2014). In the third experiment, we tested whether differences in reaction times and/or confidence due to manipulating the design of the monitoring task (requiring explicit retrieval) would result in consequences for subsequent control.

The Present Study

The overall purpose of this study was to detect whether age differences would arise when monitoring occurs within the context of retrieval. Further, we were interested in whether the influence of these factors and age would change when the cognitive demands surrounding the selection of a subset of items were manipulated.

Predictions and analyses. We predicted a negative relationship between retrieval success (when measured) and item selection (selecting a given item for restudy) as well as a negative relationship between monitoring judgments and item selection. We examined the Goodman-Kruskal gamma (*G*) correlation between monitoring and selection decisions (Benjamin & Diaz, 2008) in order to examine *relative* strategy implementation. This allowed us to test whether a given subject chose to restudy items she rated with lower relative values of confidence, regardless of the absolute magnitude (mean) confidence.

We expected that the strength of these relationships would differ for young and older adults (Souchay & Isingrini, 2004). Examining both a relative measure (gamma) and magnitude (the specific numerical value on the 0 – 100 scale) enabled us to determine whether young and older adults differed in relative accuracy, in magnitude, in both, or in neither.

Additional analyses. We also examined factors that affect or could affect monitoring and the relationship between those factors and control. Specifically, we examined valence and partial information. First, target items used within these studies had either a positive or negative valence (which in the design served as the basis for contextual/partial information about the to-be-remembered word). We evaluated whether

target valence was related to monitoring and control. Second, we examined partial information (when individuals were not able to successfully retrieve the target but were able to retrieve its valence) and its relationship to monitoring and control.

Alternatives to by-participant analysis. Finally, researchers have suggested limitations of the use of gamma in metamemory research (Murayama, Sakaki, Yan, & Smith, 2014; Masson & Rotello, 2009; Benjamin & Diaz, 2008). Most importantly, the use of gamma to capture group-level differences leads to a greater possibility of Type I error when random effects are present. More specifically, traditional by-participant analysis considers only the random variation that exists across participants, and no other random variation, such as the variation that exists across items. Items included within an experiment are ideally a random sample from a population; however, the random sampling variation of items likely results in random variation in the outcome variable across items. When not accounted for, this random item variation can increase the likelihood of finding a significant effect where there is none (see Murayama et al, 2014 p. 1290 – 1291 for complete discussion). Mixed-effects modeling, which accounts for variation across participants and across items, has been offered as an alternative approach (Murayama et al., 2014). Therefore, we investigated the relationship between monitoring and control via this method. Broadly, our plans for model building were this: to include relevant predictor variables (as fixed effects) and to model these in the presence of random variation of participants and items. If significant results were found in our proposed by-participant analyses, we could add further confirmatory evidence (as well as examine the influence of multiple predictors in one model) via mixed-effects modeling. Model specification and procedures are described in forthcoming sections.

Experiment 1

The objective of the first study was to test whether the relationship between FOK and control was different for young and older adults. In this study monitoring occurred within the context of an explicit retrieval attempt; following monitoring, control was implemented as the decision to restudy a subset of items.

Method

Participants

We identified a total sample size of 128 (4 groups of 32) via statistical power analysis using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). We specified an estimate of a medium effect size *f* (.25, Cohen, 1988), a standard α error probability (.05), and estimated power (1-β) of .80. This sample size estimate was conducted based on the anticipated use of the ANOVA procedure to address group differences in average within-subject correlations between FOK and the decision to restudy.

We tested young adults from Tufts University and older adults from the Greater Boston area. Sixty-four undergraduate and graduate students (30 males and 34 females), ranging in age from 18–27 (*M* = 19.06, *SD* = 1.4), from Tufts University participated for partial fulfillment of class requirements. Sixty-four community-dwelling older adults (18 males and 46 females), ranging in age from 57–87 years (*M* = 71.25, *SD* = 6.7), participated in exchange for nominal compensation ($15). All older adults were administered the National Adult Reading Test (NART; Nelson & Willison, 1991), validated for American older adults (Grober, Sliwinsk, & Korey, 1991). All older adults had an estimated premorbid IQ greater than 100 (which ranged from 103.84 to 131.44). No participant was excluded on the basis of performance on the NART.

Older adults had higher mean years of education, $M = 16.90$, $SD = 2.6$, compared to younger adults, $M = 13.07$, $SD = 1.9$, $t(113.7) = 9.44$, $p < .001$, as well as higher scores on the Shipley vocabulary scale (Zachary, 1991), $M = 15.88$, $SD = 2.2$, relative to younger adults, $M = 13.66$, $SD = 1.7$, $t(126) = 6.30$, $p < .001$.

Materials

The items utilized in this experiment consisted of 36 word pairs with low forward and backward association strength, from Thomas, et al. (2011). Target words had either a positive or negative valence (Warriner, Kuperman, Brysbaert, 2013), and were matched for word frequency with the respective cue words.

Procedure

Participants provided informed consent prior to participating in the experiment, which was conducted with E-Prime software (Psychology Software Tools, Pittsburgh, PA) on Windows personal computers. Young and older participants were randomly assigned to the between-subjects manipulation, employed by Thomas et al. (2011), based on their order of arrival at the lab. Specifically, during the initial testing phase, participants received one of two question orders: valence first or FOK first. In the valence first condition, participants were first asked about the valence of the target word prior to making the FOK judgment, which has been shown to increase older adults' FOK accuracy and magnitude in episodic FOK tasks (Thomas et al., 2011). The remaining half of the participants received the two questions in the opposite order, with the valence question second.

Some differences in testing conditions for the age groups were undertaken. Older adults were tested individually, as responses were spoken out loud and typed by the

experimenter. Younger adult participants were tested in groups of one to four, and all responses were typed via keyboard by the participants. Differences in testing between young and older adults is a common practice (e.g. Thomas, et al., 2011; Marsh, Dolan, Balota, & Roediger, 2004), as older adults may have less familiarity and experience less comfort with computer based testing (American Psychological Association, 2014, p. 39). Experimenter entry of typed responses likely influenced responses times; however, comparing response times of age groups was not of interest to this or the second experiment. Further, note that there was no difference in young adult recall performance as a function of testing group size[1].

Comparing groups that have different mean levels of test performance can complicate interpretation of metacognitive results (Schwartz & Metcalfe, 1994), thus, different study times and number of word pairs were employed to try to mitigate differences between groups. A presentation rate of 4 s per pair for younger adults and 8 s per pair for older adults were selected, as different rates are commonly used (e.g. Tullis & Benjamin, 2012). Younger adults were presented with 30 word pairs and older adults were presented with 24 word pairs. All other aspects of the procedure were the same for young and older adults.

Initial study phase. Participants received word pairs in a random order during the initial study phase. Word pairs appeared in the center of the computer screen in black Garamond 32-point font on a white background. Between each word pair presentation, a blank screen appeared for 0.5 s.

[1] A one-way between-subjects ANOVA was used to test whether test performance at either testing phase (initial test and final) differed as a function of testing-group size. No differences in initial or final test performance emerged, *p*s>.42.

Initial testing phase. After studying all word pairs, participants were presented with each cue word individually and were asked to recall the target word. After the recall attempt, participants received two question prompts: one about feeling of knowing and the other about the valence of the target word. Feeling of knowing was prompted for every item (Schwartz, Boduroglu, & Tekcan, 2016) and was phrased as: *If you can study this item later, how likely do you think it is you can remember the correct answer in the future memory test?* A scale was displayed at the bottom of the screen from 0 to 100 (in increments of 10), where 0 was '*not likely*' and 100 '*very likely*'. Participants were also asked about the valence of the target word, and could respond '*no answer'*, '*positive*', or '*negative*' (keys 0, 1, 2 respectively). After participants had made their responses, the next cue was presented on the screen, and this continued until the participants made a recall attempt and answered the questions for every studied cue.

Restudy and final test. After finishing the rating phase, participants were instructed to select half of the items to restudy. The entire set of cues was presented on the screen in an array comprised of 6 columns (Thiede & Dunlosky, 1999). Four array orders were presented across participants, so that each cue word appeared an equal amount of times in the four different quadrants of the computer screen. Each cue was labeled with a number, and items were selected by typing the cue's corresponding number. The cues disappeared from the array when they were selected, and the number of items that remained to be selected was updated to display the selections remaining. After all cues had been selected, participants began the restudy phase. During the restudy phase, word pairs were re-presented similarly to the first study phase, but in a new random order. Finally, participants engaged in the cued recall task that included all

studied pairs, where they were given each cue and asked to type in the appropriate target item. Items were tested in a new random order, and the cue remained on the screen through the response. Only the initial study and restudy sessions were presented at a fixed rate; all remaining phases of the experiment (FOK, valence, cued recall, and the selection phase) were self-paced.

Results

First, we present analyses of group differences in factors influencing monitoring and control, such as recall accuracy and access to partial information. We then present group differences in FOK magnitude and within-person correlations of resolution. Finally, we present generalized multilevel regression models to examine factors that predict the likelihood of selecting an item to restudy.

Due to a programming error, some FOK data points were missing (1.3% in Experiment 1 and 2.1% in Experiment 2). We believe that the probability of the missing data was completely random, and, the trials themselves were omitted (i.e., the number of trials could vary across participants depending on whether they had complete data on FOK) from only the analyses which included FOK. Finally, in cases where the assumption of sphericity was not met, the Greenhouse-Geisser correction was applied.

Factors Influencing Monitoring and Control

Recall Accuracy. Young and older adults differed in mean test performance. A 2 (age: young, older) X 2 (question group: valence first, FOK first) factorial ANOVA was performed on mean initial test and mean final test performance. There were significant differences in initial test, $F(1, 124) = 6.15$, $p < .05$, $\eta_p^2 = .05$, and final test, $F(1, 124) = 14.90$, $p < .001$, $\eta_p^2 = .11$, performance between young and older adults.

Younger adults demonstrated higher average scores ($M_{initial}$ = .34, SD = .17, M_{final} = .65, SD = .21) than older adults ($M_{initial}$ = .26, SD = .20, M_{final} = .49, SD = .23). All other effects (question group, question group and age interaction) were not significant, Fs < 1.

To investigate the relationship between recall accuracy and the decision to restudy, we computed Goodman-Kruskal gamma (G) correlation for individuals using two dichotomous variables: initial target recall and whether the item was selected to be restudied. Two older adult participants were excluded, as G could not be computed because initial test performance did not vary. As expected, recalling the target and subsequently selecting it to be restudied were negatively related, M = -.52 (SD = .67). Interestingly, average G between young and older adults was different, $t(87) = 4.18$, $p < .001$, M_{young} = -.76 (SD = .39) and M_{older} = -.28 (SD = .81). The average relationship between initial target recall and selection was weaker for older adults.

Recall accuracy was also related to FOK, M = .86 (SD = .30). In this instance, average G between young and older adults was not different, $t(70) = 1.90$, $p = .06$. Both age groups demonstrated a positive relationship: initially recalling the target was associated with higher relative ratings of FOK.

Partial information. We tested group differences in access to partial information and in how FOK varied between items for which valence (partial information) was accurately retrieved versus not. These analyses were conducted only on those items not recalled correctly at initial test. A 2 (age group: young, older) X 2 (question group: valence first, FOK first) between-subjects ANOVA was performed on the mean

proportion[2] of items for which partial information was successfully retrieved. No group differences emerged, *F*s < 1. Participant groups accessed partial information at similar rates, ranging from 30 to 33 percent of the time.

A 2 (partial information: correct, incorrect) X 2 (age group: young, older) X 2 (question group: valence first, FOK first) mixed design ANOVA was performed on mean FOK. There was a main effect of partial information, $F(1, 117) = 179.66$, $p < .001$, $\eta_p^2 = .61$. When individuals did not correctly recall an item, but did correctly recall its valence, FOK was higher than when they did not correctly recall its valence, $M_{correct} = 48.98$, $SD = 20.0$; $M_{incorrect} = 31.46$, $SD = 21.4$.

Further, there was a significant interaction between partial information and age group, $F(1, 117) = 10.09$, $p < .01$, $\eta_p^2 = .08$. Young adults demonstrated higher FOK when valence was correctly recalled as compared to older adults (M_{young}=53.58, M_{older}=44.30); however, there was no difference between older and younger adults when valence was not correctly recalled (M_{young}=31.97, M_{older}=30.96). Neither the effect of question group, $F(1, 117) = 2.76$, $p = .10$, nor its interaction with partial information, $F(1, 117) = 2.93$, $p = .09$, was significant.

Finally, we investigated the relationship between partial information and the decision to restudy, via average within-subject correlations. Fourteen participants were excluded (6 young adults, 8 older adults), as *G* could not be computed because either partial information or the selection of items to be restudied was constant. Accessing partial information and subsequently selecting it to be restudied were negatively related,

[2] Young and older adults received a different number of items to learn, therefore proportions were calculated for each individual such that the number of items for which partial information was successfully retrieved was divided by the number of possible items (items incorrect at initial test).

but only for young adults, $t(112) = 2.14$, $p < .05$, $M_{young} = -.22$ ($SD = .74$) and $M_{older} = .06$ ($SD = .64$). For young adults, the average relationship was significantly different than zero, $t(57) = 2.27$, $p < .05$; whereas for older adults, it was not, $t(55) = 0.67$, $p = .51$.

Valence. We examined differences between targets with positive and negative valence in test performance, FOK, and item selection. A 2 (age: young, older) X 2 (valence: positive, negative) mixed-design ANOVA was performed on mean final test performance. Final test performance differed by valence, $F(1, 126) = 30.62$, $p < .001$, $\eta_p^2 = .20$. Positive targets were remembered at a greater rate compared to negative targets, $M_{positive} = 60.95$ ($SD = 24.0$), $M_{negative} = 52.11$ ($SD = 25.3$). While there was a significant effect of age, $F(1, 126) = 13.62$, $p < .001$, as older adults performed worse than younger adults, the interaction between valence and age group was not significant, $F(1, 126) = 0.74$, $p = .39$.

There was a corresponding association between valence and mean FOK, as cues with positive target valence were accorded higher average FOK than cues with negative target valence, $M_{positive} = 50.96$ ($SD = 21.4$), $M_{negative} = 46.31$ ($SD = 20.0$), $F(1,126) = 28.80$, $p < .001$, $\eta_p^2 = .19$. Although older adults' average FOK was different than that of younger adults, $F(1, 126) = 10.94$, $p < .01$, $\eta_p^2 = .08$, the interaction between valence and age group was not significant, $F(1, 126) = 2.26$, $p = .14$. Both age groups afforded higher FOK to word pairs with positive targets; both age groups also recalled more positive targets.

We computed *G* correlations for individuals using target valence and whether the item was selected to be restudied. Target valence was not related to selecting an item to be restudied, $M = .05$ ($SD = .39$), as this average relationship was not statistically

different than zero, $t(127) = 1.60$, $p = .11$. Note that no age differences were present, $t(126) = 1.24$, $p = .22$, $M_{young} = .10$ ($SD = .34$) and $M_{older} = .01$ ($SD = .43$).

Selection of Items for Restudy

FOK magnitude. Average FOK was subjected to a 2 (age group: young, old) X 2 (question group: valence first, FOK first) X 2 (item selection: selected, not selected) mixed design ANOVA, with item selection as the within-subjects factor. There was a significant difference between average FOK for items selected and not selected for restudy, $F(1,124) = 82.92$, $p < .001$, $\eta_p^2 = .40$, as well as a significant difference in average FOK between young and older adults, $F(1,124) = 11.01$, $p < .01$, $\eta_p^2 = .08$. These main effects were qualified by a significant interaction between item selection and age, $F(1,124) = 22.28$, $p < .001$, $\eta_p^2 = .15$. While young and older adults displayed similar average FOKs for items selected for restudy, $t(126) = 1.03$, $p = .31$, ($M_{young} = 42.9$, $SD = 18.7$, $M_{older} = 39.4$, $SD = 20.3$), there was a significant difference between age groups in average FOK for items not selected for restudy, ($M_{young} = 65.6$, $SD = 20.9$; $M_{older} = 46.6$, $SD = 25.5$), $t(126) = 4.63$, $p < .001$. Examination of Figure 1 further illustrates this difference.

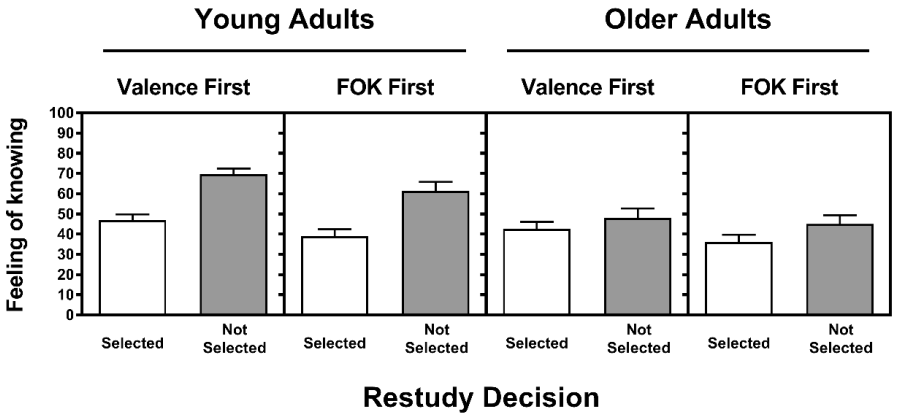

Figure 1. Average feeling of knowing rating by item selection, for young and older adults in two question groups in Experiment 1.

For younger adults, average FOK of items selected for restudy was lower than average FOK of items not selected. Older adults did not exhibit that same degree of difference, as average FOK was more similar between items selected and not selected for restudy. Differences in average FOK between the question groups were not significant, ($F(1,124) = 3.61$, $p = .060$), nor were their interactions (Fs<1).

Intra-individual correlations. To assess the relative relationship between monitoring judgments and item selection, and to facilitate comparisons to other studies, we calculated gamma correlations between FOK and whether the item was selected to be restudied. Average correlations were subjected to a 2 (age group: young, older) X 2 (question group: valence first, FOK first) factorial ANOVA, to assess whether there were age differences in the degree of the relationship between FOK and item selection. A significant main effect of age was found, $F(1,124) = 13.17$, $p < .001$, $\eta_p^2 = .10$. Neither the effect of question group nor the interaction term (age by group) was significant, Fs<1.

Average gammas are reported in Table 1. All participant groups demonstrated an average gamma correlation that was negative. Young and older adults, on average, selected items for restudy that they had given lower relative subjective ratings to. While both age groups exhibited negative average *G*, this average relationship was more pronounced for young adults, $M_{young} = -.55$ ($SD = .38$), $M_{older} = -.23$ ($SD = .59$).

Mixed-Effects Model

We conducted generalized multilevel regressions to predict the likelihood of selecting an item to restudy, using R (R Core Team, 2013). We elected to separately model items by whether they had been successfully recalled at the initial test. We hypothesized that the recollective experience (feeling of knowing and access to partial

information) is particularly relevant in cases where the target has not been successfully retrieved. However, feeling of knowing may be relevant in cases of successful retrieval (Schwartz et al., 2016).

Model specification and variable manipulation. The purpose of this mixed-effects binary logistic regression was to model the likelihood of selection with predictors at two levels. At the lowest level (Level 1) were trials; predictors at this level captured variability *within* subjects. At the highest level (Level 2) were subjects; predictors at this level captured variability *between* subjects.

Two predictors were included at Level 1: partial information and FOK. Partial information was entered into the model for unrecalled items only, and was dummy-coded such that '1' represented accurately accessed partial information. At this level, FOK was scaled into units of 10% (i.e., a value of '1' represents 10%), a type of scaling commonly done (Murayama et al., 2014). At Level 1, FOK was centered within subjects (for every trial of a given subject, that subject's mean FOK was subtracted from their trial FOK).

Two predictors were included at Level 2: age group and FOK. Here, FOK was centered between subjects (after also being transformed into units of 10%). That is, the mean FOK of participants in the sample was subtracted from each subject's mean FOK. We centered FOK at the two levels so that they acted independently, and could be used within the same model. To evaluate age differences, we dummy-coded age group such that young adults were the referent (i.e. young adults = 0, older adults = 1). Question group was not added to the models, as comparisons had revealed no significant differences between participants in the valence first and the FOK first conditions.

Model for unrecalled items. The following fixed effects were added to the model: FOK at Level 1 (within participant), partial information, FOK at Level 2 (between participant), age group, and the interaction between age group and FOK (at both levels) and age group and partial information. Random effects were included at the item and participant levels, for this and all following models (Baayen, Davidson, & Bates, 2008). As a final step, nonsignificant coefficients were trimmed. Specifically, in the model for unrecalled items, we trimmed the variables related to partial information.

To support this decision, we compared the Bayesian Information Criteria (BIC), which is an index of model fit that penalizes non-parsimony. A lower BIC is better, and a decrease in BIC larger than 2 (for each additional parameter) is considered positive evidence for model selection (Raftery, 1995). Here, the more parsimonious of our models had a lower BIC, $\Delta BIC_{df=2} = 8.6$. The included variables and the estimates of fixed and random effects may be found in Table 2.

FOK was a significant predictor of likelihood of restudy at both levels. For young adults (the reference group used), a unit increase in FOK at the trial level reduced the log-odds of selection by .24. That is, if we compare two items that differ in one unit of FOK (10%), we can expect the odds of selection for the higher FOK to be .79 that of the odds of selecting an item with one unit lower in FOK (for young adults). The interaction between within-subject FOK and age was significant, $z_{age*FOK} = 3.30$, $p < .001$. The significant interaction indicates that the relationship between FOK and the likelihood of restudy was different for older adults relative to younger adults.

In Figure 2, one can examine how the predicted probability of selecting an item to be restudied varied across units of FOK. Further, this figure illustrates how the predicted probability decreases more steeply across units of FOK for young adults. Older adults did not demonstrate as strong of a relationship, mirroring the results from the intra-individual correlations.

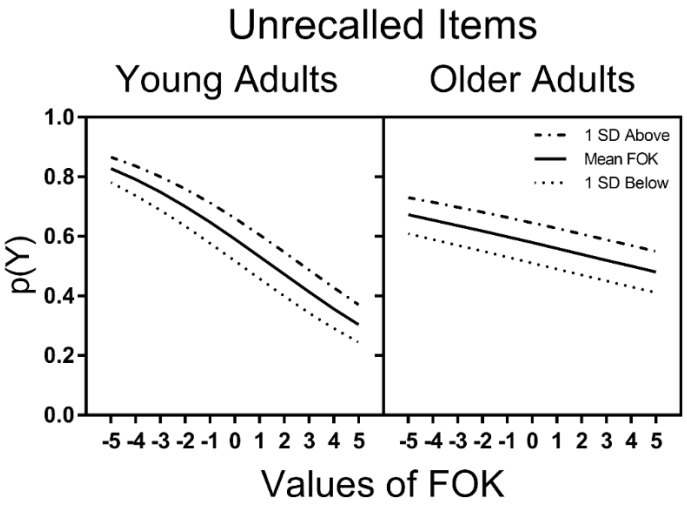

Figure 2. Predicted probability of selecting at item to be restudied, across units of Level 1 FOK, plotted separately for young and older adults. Separate lines indicate different Level 2 FOK values.

Finally, the average number of items selected to be restudied varied by between-subjects FOK. Comparing two individuals that differ in one unit of mean FOK (10%), we can expect the selection rate for the person with the higher FOK to be 1.16 that of the person with the lower FOK. That is, people with higher FOK have greater odds of selection, as can be seen when converted to units of probability in Figure 2. The effect of average FOK did not differ between young and older adults. As an illustration, the change in average FOK (plotted as 1 standard deviation above and below the mean) influences young and older adults similarly.

Model for correctly recalled items. When items were initially recalled, there was large variance between people in log odds of selection of items for restudy (see Table 2). Young adults were approximately half as likely (exp(β)= .49) to select an item for restudy as compared to not selecting an item for restudy, holding all other variables constant. Age was associated with an increase in the log-odds of selection (.57). For older adults, the odds an item was selected for restudy was 1.76 times the odds for young adults; however, because of the large standard error, this trend cannot be considered reliable.

A unit increase in trial FOK reduced the log-odds of selection (by .35), indicating that as FOK increased, young adults were less likely to select an item for restudy. There was not a significant age by FOK interaction, $z = 0.84$, $p = .40$, indicating that the relationship between trial level FOK and selection was not different for the age groups.

Finally, the average number of items selected to be restudied varied by between-subjects FOK. Comparing two individuals that differ in one unit of mean FOK (10%), we can expect the selection rate for the person with the higher FOK to be .70 that of the person with the lower FOK. That is, for cases where the item had been successfully retrieved, people with higher average FOK had lower odds of selecting an item to be restudied. This pattern did not differ between young and older adults.

Discussion

In Experiment 1, we found age differences in selection behaviors which were largely attributable to recall accuracy and FOK. For both initially unrecalled and recalled items, higher trial FOKs were associated with a reduced probability of selecting an item to be restudied. For items initially recalled, there were no age differences in the

relationship between trial FOK and selection behavior, though older adults appeared more likely to select items they had initially recalled compared to younger adults. For unrecalled items, FOKs were associated with a reduced probability of selecting an item for restudy, but this association varied in strength between the age groups. Older adults demonstrated a weaker relationship between trial FOK and item selections.

In this experiment, we also tested whether partial information was important for self-regulated learning. At the outset, differences in the relationship between FOK and partial information were observed between the age groups, though the groups accessed partial information with similar frequency. Interestingly, partial information was related to the selection of items to be restudied for young adults (as examined via average gamma correlation). However, when added to the mixed effects multilevel model, partial information did not predict the likelihood of restudy beyond that of feeling of knowing.

Between-subjects FOK was related to the selection of items for both young and older adults. Individuals with higher average FOK were more likely than those with lower average FOK to select unrecalled items (and less likely to select recalled items). This finding suggests participants varied in strategy adoption across different values of average FOK. Those who had more confidence in the future recallability of the targets were more likely to select unrecalled items to be restudied. Those who had less confidence were more likely to select items they had previously correctly recalled. For both young and older adults, selection was dependent on mean level of confidence.

Experiment 2

In Experiment 1, similarities in self-regulated learning were found for young and older adults. These results suggest that age-related differences in control previously found may depend upon the success of prior retrieval, and trial-level FOK. Age-related differences in control were not found when we examined between-subjects FOK. Importantly, strategy adoption (between individuals with varying levels of average FOK) likely emerged as the structure of the control task fostered goal-orientation. In Experiment 2, we investigated whether these similarities would persist when control decisions were more cognitively demanding.

Specifically, the relationship between monitoring and control may be different when items are chosen for restudy from an array (Tullis & Benjamin, 2012; Theide & Dunlosky, 1999) as compared to when items are chosen sequentially (Hanczakowski et al., 2014; Dunlosky & Hertzog, 1997). To illustrate, with sequential presentation of items, executing one's goals requires that several simultaneous, internal processes are active (such as maintaining the number of items already selected in addition to considering how well each individual item has been learned) during the decision process. This added burden of keeping simultaneous goals active may be particularly detrimental for older adults' self-regulatory behavior.

The objective of the second study was to test whether cognitive demands influence the relationship between FOK and control differently for young and older adults. In contrast to the first study, where participants selected items to restudy only after all items were initially tested, participants in this experiment chose whether to restudy items after each recall attempt.

Young and older adults underwent the same procedure as the first experiment, except for two important changes. First, all participants received the valence question prior to providing FOK assessments, as there was no statistical difference in FOK magnitude when question order was reversed. Second, a restudy prompt immediately followed the FOK prompt.

Method

Participants

Thirty-two undergraduate students (15 males), ranging in age from 18–21 (M = 18.63, SD = 0.8), from Tufts University participated for partial fulfillment of class requirements. Thirty-two community-dwelling older adults (7 males), ranging in age from 62–93 years (M = 74.09, SD = 9.1), participated in exchange for nominal compensation ($15). As in Experiment 1, older adults were administered the NART, and none were excluded on this basis. Each had an estimated premorbid IQ greater than 100 (which ranged from 108.24 to 127.76). Older adults had higher mean years of education, M = 17.03, SD = 3.1, compared to younger adults, M = 12.56, SD = 0.8, $t(35.4) = 7.81$, $p < .001$, as well as higher scores on the vocabulary measure (Zachary, 1991), M = 15.71, SD = 2.6, relative to younger adults, M = 13.97, SD = 1.7, $t(51.8) = 3.12$, $p < .01$.

Procedure

All aspects of the procedure were kept consistent with Experiment 1. The only change was that after participants were asked to rate FOK, a new screen appeared which displayed only the cue and asked, *"Would you like to restudy this word pair?*", to which the participant could respond '*Yes'* or '*No*' (keys 1, 2 respectively). Participants were limited to choose a subset of the items to restudy (exactly half), as has been done in prior

studies (i.e. Dunlosky & Hertzog, 1997). Two counters appeared on the screen: one that indicated the trial number (i.e. *"Trial 2 of 24"*) and one that indicated how many choices the participant had left. Once a participant had selected the allotted number of choices, the restudy prompt no longer appeared.

Results

To assess whether monitoring judgments were associated with the decision to restudy items, gamma correlations were computed between FOK and item selection decisions. Note that *G* could not be computed for 1 young and 1 older adult as FOK did not vary. Visual examination of Table 1 demonstrates that mean within-subject correlations were negative. In accordance with findings from Experiment 1, participants were more likely to select an item for restudy that they had assigned a lower FOK value. There was a significant age difference, t(60) = 2.26, p = .028.

For young adults, average gamma correlations for item selection decisions were significantly different than zero, (M = -.48, SE = .08), $t(30) = 5.93$, $p < .001$, however, older adults' gamma did not significantly differ from zero, (M = -.17, SE = .11), $t(30) = 1.60$, $p = .12$, lending evidence to our hypothesis that the cognitive demand of sequential control decisions at initial test resulted in a weakened association between FOK and control for older adults.

Factors Influencing Monitoring and Control

Recall Accuracy. A t-test was performed on mean initial test performance, and revealed a significant difference in initial test performance between young and older adults $t(62) = 2.79$, $p < .01$, 95% CI [.03, .21]. Younger adults performed better on the initial test, $M_{younger}$ = .34 (SD = .18), M_{older} = .23 (SD = .16). A t-test was performed on

mean final test performance, and revealed a significant difference in final test performance between young and older adults $t(62) = 3.79$, $p < .001$, 95% CI [.09, .30], $M_{younger} = .65$ ($SD = .21$), $M_{older} = .45$ ($SD = .20$).

Partial information. A t-test was performed on the mean proportion of items for which partial information was successfully retrieved. No age differences were found, $t(62) = 0.58$, $p = .57$, $M_{younger} = .38$ ($SD = .21$), $M_{older} = .35$ ($SD = .21$). A 2 (partial information: correct, incorrect) X 2 (age group: young, older) mixed design ANOVA was performed on mean FOK. Partial information was significant, $F(1, 59) = 62.06$, $p < .001$, $\eta_p^2 = .51$. When individuals did not correctly recall an item, but did correctly recall its valence, FOK was higher than when they did not correctly recall its valence, $M_{correct} = 53.60$, $SD = 20.8$; $M_{incorrect} = 39.53$, $SD = 19.4$.

While there was a main effect of age, $F(1, 59) = 5.49$, $p < .05$, $\eta_p^2 = .09$, age did not interact with partial information, $F(1, 59) = 1.65$, $p = .21$. That is, successfully retrieving partial information affected average FOK similarly for young and older adults alike. Partial information was entered into the model predicting likelihood of restudy, as in Experiment 1.

Valence. A 2 (age: young, older) X 2 (valence: positive, negative) mixed-design ANOVA was performed on mean final test performance and on mean FOK. There was no significant effect of valence, $F(1, 62) = 1.94$, $p = .17$, or a significant interaction between valence and age, $F(1, 62) = 0.24$, $p = .62$, on test performance. Targets with positive valence and targets with negative valence were recalled at equal rates. However, items with positive target valence were accorded higher average FOK than items with negative target valence, $M_{positive} = 54.71$ ($SD = 19.4$), $M_{negative} = 51.64$ ($SD = 19.8$),

$F(1,62) = 8.01$, $p < .01$, $\eta_p^2 = .11$. While older adults' average FOK was different than that of younger adults, $F(1, 62) = 8.60$, $p < .01$, $\eta_p^2 = .12$, the interaction between valence and age group was not significant, $F < 1$.

Mixed-Effects Model

The same model procedure was followed for Experiment 2 with one exception. Cases for which the restudy prompt was never asked were omitted (these cases could never have been selected so the likelihood of selection was known, P(Y) = 0). Unlike in Experiment 1, between subjects FOK was trimmed from the models (the addition of the two parameters resulted in an *increase* in BIC, $\Delta BIC_{unrecalled} = 12.4$, $\Delta BIC_{recalled} = 11.1$).

Model for unrecalled items. Model results are reported in Table 2. When items were not recalled at initial test, young adults were approximately 6 times as likely to select an item for restudy as compared to not selecting an item for restudy. Age was significant, $z = -2.08$, $p < .05$. For older adults, the odds an item was selected for restudy was roughly half the odds for young adults, holding all other variables constant (including random effects). Level 1 FOK and the interaction of FOK and age were not significant, zs < 1, both exhibiting odds ratios close to 1. FOK had no impact on the log-odds of selection for either young or older adults.

For young adults, successfully retrieving partial information was associated with a reduced log-odds of selection (1.19). For young adults, the odds of selection for an item where partial information had been retrieved were roughly a third (.31) that of the odds of selecting an item where partial information had not been retrieved. The interaction between partial information and age was significant, $z = -4.83$, $p < .001$.

Figure 3 illustrates how the predicted probability differed by access to partial information for young adults, but not for older adults.

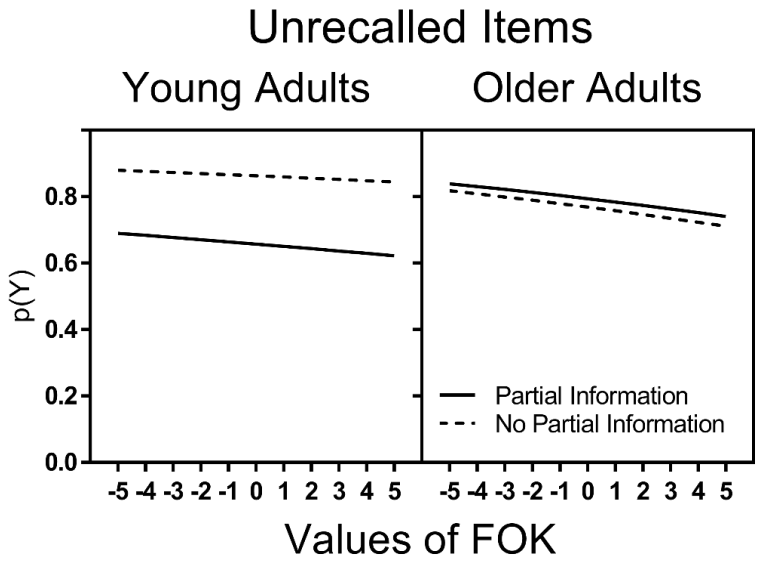

Figure 3. Predicted probability of selecting at item to be restudied, across units of Level 1 FOK, plotted separately for young and older adults. Separate lines indicate success of retrieved partial information.

Model for correctly recalled items. When items were initially recalled, there was large variance between people in log odds of selection of items for restudy, as in Experiment 1. Age was not significant, $z = 1.51$, $p = .13$. The estimate was close to that of Experiment 1. For items that were initially recalled, older adults' odds of selecting an item to restudy was close to three times (2.97) that of younger adults. However, as indicated by the large standard error, this trend cannot be considered reliable.

Unlike incorrect items, a unit increase in FOK significantly reduced the log-odds of selection, indicating that as FOK increased, young adults were less likely to select an item for restudy. Specifically, if two items differ in one unit of FOK (10%), we would expect the odds of selection for the higher FOK to be roughly a third that of the odds of selecting an item with one unit lower in FOK. The interaction between age and FOK was

not significant, z = - 0.39, p = .23, older adults did not differ from young adults in the relationship between FOK and selection for recalled items.

Discussion

Results from Experiment 2 differed from those of Experiment 1 in a few notable ways. First, the strength of the relationship between FOK and restudy (as measured by gamma correlations) was reduced. For older adults, that relationship was not statistically different from zero. However, results from the mixed-effects models diverged from those of the by-participant analysis. For unrecalled items, neither within-subject nor between-subject FOK predicted the likelihood of restudy. Within-subject FOK was a significant predictor only when items had been successfully recalled at the initial test. Interestingly, accessing partial information emerged as a significant predictor of the likelihood of restudy. When young adults successfully accessed partial information about an unrecalled target, they were less likely to select that item to be restudied. This effect was not present for older adults, who were similarly likely to select an item to be restudied whether they had successfully retrieved partial information or not.

Experiment 3

In Experiment 1, young adults demonstrated a stronger relationship between monitoring and control than older adults. In Experiment 2, that pattern was even stronger. The primary difference between these two studies was that in Experiment 2 the decision to select items was made sequentially at test as opposed to presenting all items simultaneously for selection. Increasing the demands of the selection decision (control) negatively impacted the relationship between monitoring and control for older adults particularly (the relationship was not statistically different than zero), however,

it remains unknown whether increasing the demands of the monitoring phase would result in a similar pattern.

In the first two experiments, participants in all groups had to go through the same exact three stages: explicit retrieval, FOK question, and valence question. All manipulations likely promoted explicit retrieval.

There is a related literature on explicit retrieval which employs judgments of learning (JOL) known as the pre-judgment recall and monitoring methodology (PRAM; Nelson, Narens, & Dunlosky, 2004). This procedure, in which an initial recall stage is followed by a monitoring judgment stage, is different than traditional JOL methodology that combines the two, where retrieval is covert and concurrent with the judgment. In their study, while judgments made after a delay had higher relative accuracy than judgments made immediately after study presentation (Nelson & Dunlosky, 1991), PRAM-methodology delayed-JOLs and traditional delayed-JOLs exhibited equivalent relative accuracy.

Despite equivalence in the accuracy of monitoring between methodologies, it is possible that cognitive effort differed. Son and Metcalfe (2005) found that reaction times to make traditional delayed-JOLs differed from reaction times when people were told to first attempt retrieval (either overtly or covertly) and then make JOLs. The traditional delayed-JOLs, where retrieval was not explicitly instructed, exhibited inverted U shape reaction time functions: the time spent on making judgments for items that received the lowest and highest ratings was much less than the time spent making judgments for items that received middle-range ratings (i.e. 30 – 60 on a 100-point scale). When retrieval was explicitly instructed, however, the retrieval attempt reaction time demonstrated a linear

relationship with JOL. Reaction times for the lowest rated items were longest, and decreased monotonically with increasing JOL rating.

These findings suggest that cognitive effort may differ when participants make an explicit retrieval attempt before making JOLs as compared to when they do not. The difference in effort may have downstream consequences on restudy decisions, which may differentially impact older as compared to younger adults. Specifically, older adults, who demonstrate a decline in available cognitive resources (Salthouse, 2010), may be differentially affected when retrieval demands are increased. The purpose of Experiment 3 was to investigate age differences in the relationship between monitoring and item selection decisions, when retrieval was explicitly instructed or not explicitly instructed.

Method

Design

This experiment was a 2 (age group: young, old) X 2 (monitoring: Retrieval + JOL, JOL only) between-subjects factorial design. As in Experiment 1, participants were randomly assigned to the between-subjects manipulation based on the order of arrival at the lab. Sample size was determined via the same procedure as Experiment 1; however, we used our data from study 1 to estimate effect size. Further, though effect sizes may be computed for each effect and interaction, our estimate was based on our observed effect of age on the relationship between monitoring and control ($\eta_p^2 = .10$ or $f = .33$). With these parameters ($f = .33$, $\alpha = .05$, $1-\beta) = .80$), our power calculation yielded a total sample size of 75, which was rounded up to 108 in order to allocate participants to groups and running orders (counterbalancing) evenly.

Participants

We recruited young adults from Tufts University and older adults from the Greater Boston area. Fifty-four undergraduate students (23 males), ranging in age from 18–21 (*M* = 18.9, *SD* = 0.8), from Tufts University participated for partial fulfillment of class requirements. Fifty-four community-dwelling older adults (15 males), ranging in age from 58–79 years (*M* = 69.9, *SD* = 4.8), participated in exchange for nominal compensation. Older adults had higher mean years of education, *M* = 17.24, *SD* = 2.5, compared to younger adults, *M* = 12.57, *SD* = 0.7, $t(62.2) = 13.1$, $p < .001$.

The remaining pool of eligible older adult participants was small. Previously, young and older adult participants were not eligible for Experiment 2 if they had participated in Experiment 1 (as these studies used the same stimuli). Consequently, we developed new word pairs for this experiment so that older adults who participated in the two previous experiments were not excluded from participating in the third. Of the participants included in Experiment 3, twenty-three had participated in Experiment 1 ($N_{Retrieval + JOL} = 10$; $N_{JOL\ only} = 13$), thirteen had participated in Experiment 2 ($N_{Retrieval + JOL} = 9$; $N_{JOL\ only} = 4$)[3]. Eligibility for younger adult participants remained the same.

Materials

Words were randomly selected and assigned to cue-target pairs, which were then checked for semantic relatedness and frequency of co-occurrence. Seventy-two words were selected from the MRC Psycholinguistics Database (Coltheart, 1981) to be displayed as cues and targets. As in the previous studies, the cues had a neutral valence

[3] A one-way between-subjects ANOVA was used to test whether performance (study selection behavior and monitoring ratings) differed as a function of previously participating in Experiments 1 or 2. No differences emerged, *p*s>.46.

and the targets had either a positive or negative valence (Warriner, Kuperman, Brysbaert, 2013). Using the University of South Florida Free Association Norms (USF FAN; Nelson, McEvoy, & Schreiber, 1998), we checked that no target word occurred in the set of possible associates for its cue, and vice versa. We did this to help ensure that all paired words were of the same approximate difficulty (for example, cue-target pairs that are associates of one another are easier to remember). Finally, a pairwise latent semantic analysis (LSA; lsa.colorado.edu, Landaver, Foltz, Laham, 1998) was used to examine the extent to which the paired words co-occur. We did this to help ensure that none of the words paired co-occur with a significantly different rate of frequency than the others.

Procedure

As in the previous experiments, we chose to differ the study time and number of words between age groups. A presentation rate of 2.5 s per pair for younger adults and 7.5 s per pair for older adults was used. Younger adults were presented with 30 word pairs and older adults were presented with 24 word pairs. All other aspects of the procedure were the same between young and older adults, as all underwent the initial study phase, monitoring phase, restudy, and final test, and were tested individually. The initial study phase and the final test were identical to the previous experiments; differences in the other phases are denoted below.

Monitoring phase. After studying all word pairs, participants were presented with each cue word individually. During this stage, both young and older adults made responses aloud at the same time as pressing the spacebar, so that reaction times could be collected. A microphone recorded the audio, and the experimenter transcribed the answers after the study concluded.

Every participant made a JOL. That is, the cue (the first word of the pair) was presented on the screen, at which time the following question appeared at the top of the screen: *"How likely do you think it is you can remember the correct answer in the future memory test?"* This differed from Experiments 1 and 2 as we omitted the phrase, *"If you can study this item later"* as a part of the monitoring prompt, as not all words were included in the monitoring phase (see below). The same 0 to 100 scale was used as in the previous experiments. However, instead of typing their responses, participants reported the number out loud (at the same time as pressing the space bar), after which the next cue word was presented.

Participants within the Retrieval + JOL condition viewed an additional screen prior to each JOL, where they were first asked to try to recall the target word (the second word of the word pair). During this retrieval stage, the following prompt appeared above the cue: *What was the second word?* Participants said the answer out loud (or said "next" if unable to recall it) while pressing the space bar to proceed to the monitoring stage. Thus, participants within the Retrieval + JOL condition responded to two screens, the retrieval and JOL screens, one after the other. The JOL only group received only the JOL screen. All participants had unlimited time to make their responses (Son & Metcalfe, 2005). Two reaction times were recorded for participants in the Retrieval + JOL group (the button-press to the two screens), while one reaction time was recorded for participants in the JOL only group (the button-press to the JOL screen).

Restudy and final test. After participants had rated all items, a new screen appeared which asked, *"How many word pairs would you like to restudy?"* Because participants were instructed to select a finite number of cues to restudy (as in the first two

experiments), this question was asked to ascertain how many cues participants would select if able to so without constraint.

After this screen, participants were told to select half of the items to restudy from a subset of cues that were presented on the screen in an array comprised of 4 rows (Thiede & Dunlosky, 1999). One-third of the cues were excluded from the restudy phase to measure monitoring prediction accuracy in the absence of the choice to restudy. Excluded cues were counterbalanced across three subsets, so that each cue word set was excluded from restudy an equal amount of times. Three array orders were presented across each of the three subsets, so that each cue word appeared an equal amount of times in three sections of the computer screen. As in Experiment 1, each cue was labeled with a number, and items were selected by typing the cue's corresponding number. Participants selected half of the displayed cues (8 for older adults and 10 for younger adults). After all cues had been selected, participants began the restudy phase, followed by the final test.

Results

Factors Influencing Monitoring and Control

Reaction times. We conducted generalized multilevel models to predict reaction times from judgments of learning. We separately modeled the reaction time data for participants in the between-subjects monitoring groups. Specifically, in accordance with findings from Son and Metcalfe (2005), we expected participants in the JOL only group to exhibit an inverted u-shaped relationship between JOL and reaction times. We tested this hypothesis via likelihood ratio test (LRT, Baayen et al., 2008). To calculate this, a mixed-effects model is applied to the data set. A second model is then fit to the same data set, after the addition of one or more fixed or random effects (such as the fixed effect of

trial JOL). Comparison of the fit statistics between the two models, using an LRT, signals whether there had been an improvement in model fit (as a result of the addition or subtraction of effects). If the relationship between JOL and reaction times is not linear, JOL squared will significantly reduce the deviance when added to the model.

We also employed LRT to test whether we would replicate results from the Retrieval + JOL group in Son and Metcalfe (2005). To do this, we tested hypotheses about each of the separate reaction times within this condition (retrieval button-press, JOL button-press). In accordance with findings from Son and Metcalfe (2005), we expected that the time taken to retrieve (retrieval button-press) would demonstrate a relationship where reaction times decreased linearly with increasing JOL. If this is the case within our study, adding JOL as a predictor should significantly reduce deviance to model without predictors. This would indicate a linear relationship between reaction time and JOL. Adding JOL squared should not significantly reduce deviance. If it were to reduce deviance, this would indicate a curvilinear relationship between reaction time and JOL, which would contradict the pattern of results observed in Son and Metcalfe (2005) where a strictly monotonic relationship had been observed.

Finally, Son and Metcalfe (2005) found that the reaction time of the second button-press (that to make the JOL) was a constant. Specifically, the time spent making a JOL was found to be consistent across all units of JOL. If this is the case within our study, adding JOL (and its square) should not significantly reduce deviance when added to the null model (predicting the reaction time to the second button-press, the JOL screen).

Note we did not transform RT but instead modeled raw RT in milliseconds. Reaction time data typically follows a positively skewed distribution, and thus, must be transformed to satisfy the assumption of normality. With generalized multilevel modeling, we specified an appropriate response distribution (Gamma) which best captured the distribution of our observed RT. The link function we specified was the identity link, and random effects were included for both participants and items (for discussion of the benefits of examining RT via GLMM see Lo & Andrews, 2015). Finally, reaction times larger than two and half standard deviations from an individual's mean were excluded. This resulted in the exclusion of 133 cases from the data (4.6% of possible cases).

Modeling procedure. The purpose of this generalized mixed-effects model was to include predictors at two levels. At level 1 were the trials. Predictors at this level included JOL and JOL squared. JOL was transformed into units of ten, and was lower limit centered. Thus, a JOL of zero represented the lowest possible rating (0%). The quadratic term (JOL squared) was the square of the transformed JOL. Thus, its highest possible value was 100 (10 squared). At level 2 were the subjects. Predictors at this level included age, which was dummy-coded such that young adults were the referent. A main effect of age would indicate that older adults demonstrated different reaction times than younger adults for JOLs of zero. Age could also interact with the predictors. A significant interaction would indicate that the *relationship* between the predictor (JOL) and reaction time significantly differed for older as compared to younger adults.

JOL only groups. The baseline model included the intercept and two random intercept components (item and subject). That is, reaction times were allowed to

randomly vary across items and across subjects. To this baseline model, trial JOL was added, and did not significantly reduce deviance, $\chi^2(1) = 2.16$, p = n.s. However, the addition of JOL squared and age and its interactions did significantly reduce deviance (see Table 3.). Refer to Table 4 for the summary of the complete model (interpretations discussed below).

Retrieval + JOL groups. Reaction times were analyzed separately for the button-press to the retrieval screen and the button-press to the JOL. In Son and Metcalfe (2005) the time spent making a button-press to the JOL added a constant amount of time across the values of JOL; the time spent making a button-press to complete retrieval was monotonic (decreased linearly with increasing JOL). The authors reported no curvilinear relationship between JOL and reaction time to either button-press. In our study, JOL, JOL squared, and age and its interactions significantly reduced deviance, when added to the model for the retrieval button-press *and* for the JOL button-press (see Table 3 for complete fit statistics).

Interpretations of coefficients. Refer to Table 4 for the summaries of the complete models and to Figure 4 for the predicted reaction times (in seconds) for the between-subjects manipulation. Briefly, all intercepts were statistically significant. That is, at JOL '0' young adults demonstrated reaction times that were statistically different than zero. Further, there was a main effect of age. At JOL '0' older adults demonstrated significantly greater reaction times than young adults. This difference in reaction times between the age groups can be seen clearly in Figure 4.

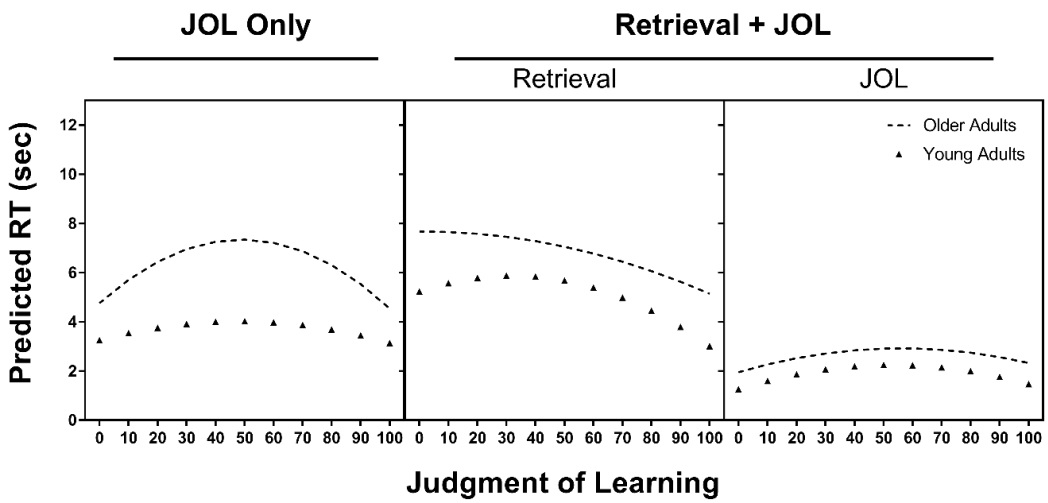

Figure 4. Predicted reaction times (in seconds) across units of Level 1 JOL plotted separately for young and older adults. Within the Retrieval + JOL group reaction times are plotted separately for the first button-press (retrieval screen) and the second button-press (JOL screen).

In the JOL only condition, young adults demonstrated a (predicted) curvilinear relationship between JOL and reaction time. There was a significant interaction with age. As can be seen in the predicted reaction times in Figure 4, older adults demonstrated a more pronounced predicted curvilinear relationship. See Figure 5 for actual RTs.

In the Retrieval + JOL condition, a curvilinear relationship was found for reaction times to the first button-press (retrieval) and second button-press (JOL). Interestingly, age differences in this relationship were found only for the button-press to retrieve. Though we did not find the monotonically decreasing relationship reported previously, an examination of predicted RTs (Figure 4) and mean RTs (Figure 5) demonstrated that there were clear differences between the participant groups. In essence, individuals in the JOL only groups demonstrated faster RTs than individuals in the Retrieval + JOL groups, and age differences were observed across these conditions.

Valence. We examined differences between targets with positive and negative valence in test performance, JOL, and item selection. A 2 (age: young, older) X 2 (monitoring: retrieval + JOL, JOL only) X 2 (valence: positive, negative) mixed-design ANOVA was performed on mean final test performance. Final test performance differed by valence, $F(1, 104) = 5.88$, $p < .05$, $\eta_p^2 = .05$. Negative targets were remembered at a greater rate compared to positive targets, $M_{positive} = 47.25$ ($SD = 26.0$), $M_{negative} = 51.09$ ($SD = 23.5$). While age was not significant, $F(1, 104) = 3.54$, $p = .063$, the interaction between valence and age group was, $F(1, 104) = 11.29$, $p < .01$, $\eta_p^2 = .10$. For young adults, accuracy was similar for targets with different valence, $M_{positive} = 54.07$ ($SD = 25.9$), $M_{negative} = 52.59$ ($SD = 25.6$). This was not the case for older adults, who exhibited better performance for words with negative targets, $M_{negative} = 49.59$ ($SD = 21.2$), compared to words with positive targets, $M_{positive} = 40.43$ ($SD = 26.0$). No other effects were significant, ps > .24.

A similar pattern was present between valence and mean JOL. Cues with negative target valence were accorded higher average JOL than cues with positive target valence, $M_{positive} = 36.57$ ($SD = 22.7$), $M_{negative} = 39.43$ ($SD = 22.1$), $F(1,104) = 5.34$, $p < .05$, $\eta_p^2 = .05$. The interaction between valence and age group was significant, $F(1, 104) = 4.19$, $p = .05$, $\eta_p^2 = .04$. The difference in average JOL between words with positive and negative valence was smaller for young adults, $M_{positive} = 40.36$ ($SD = 22.5$), $M_{negative} = 40.69$ ($SD = 21.5$), while older adults displayed a larger degree of difference, $M_{positive} = 32.78$ ($SD = 22.5$), $M_{negative} = 38.18$ ($SD = 22.8$). Finally, there was a main effect of monitoring group, $F(1,104) = 20.38$, $p < .001$, $\eta_p^2 = .16$, which did not interact with valence.

We computed *G* correlations for individuals using target valence and whether the item was selected to be restudied. A 2 (age: young, older) X 2 (monitoring: retrieval + JOL, JOL only) between-subjects ANOVA was performed on average gamma. No effects were significant. Of note, there appeared to be a trend. Individuals in the Retrieval + JOL groups demonstrated a positive gamma (M_{young} = .15, M_{older} = .16) while individuals in the JOL only groups demonstrated average gammas near zero (M_{young} = - .04, M_{older} = .02). This effect, however, cannot be considered reliable, $F(1,104) = 2.96$, $p = .051$.

Selection of Items for Restudy

Number of items. We examined differences in the number of word pairs individuals indicated they wished to restudy. A 2 (age: young, older) X 2 (monitoring: retrieval + JOL, JOL only) ANOVA was performed on the proportion of items individuals indicated they wished to restudy. Only age was significant, $F(1,104) = 4.85$, $p < .05$, $\eta_p^2 = .05$, where young adults, on average, wanted to restudy less words. No other effects were significant, *p*s > .20. However, it is of note that an opposite pattern appeared present for young and older adults. Young adults in the Retrieval + JOL group wanted to restudy about half of the items (M = 49.8%), whereas those in the JOL only group indicated they wanted to study less than half (M = 43.6%). Older adults in the Retrieval + JOL group wished to restudy just over half of the items (M = 54.6%); however, when in the JOL only group, older adults wished to restudy more items (M = 62.3%). While an interesting trend to consider, this pattern was not significant, $F(1,104) = 1.67$, $p = .20$.

JOL magnitude. Average JOL was subjected to a 2 (age group: young, old) X 2 (monitoring: retrieval + JOL, JOL only) X 2 (item selection: selected, not selected) mixed

design ANOVA, with item selection as the within-subjects factor. Note that items excluded from the restudy selection phase were omitted from this and the following analysis. There was a significant difference between average JOL for items selected and not selected for restudy, $F(1,104) = 66.80$, $p < .001$, $\eta_p^2 = .39$. While average JOL did not differ between young and older adults, $F(1,104) = 1.76$, $p = .19$, there was a significant interaction between item selection and age group, $F(1,104) = 6.53$, $p < .05$, $\eta_p^2 = .06$. Young and older adults displayed similar average JOLs for items selected for restudy, $t(106) = 0.47$, $p = .64$, ($M_{young} = 25.14$, $SD = 21.5$, $M_{older} = 27.17$, $SD = 23.0$); however, there was a difference between age groups in average JOL for items not selected for restudy, ($M_{young} = 55.64$, $SD = 29.3$; $M_{older} = 43.15$, $SD = 31.5$), $t(106) = 2.14$, $p < .05$.

The difference in average JOL between monitoring groups was significant ($F(1,104) = 20.92$, $p < .001$); interactions involving monitoring groups were not significant, Fs<1. See next page for Figure 6 for illustration of this and previously mentioned significant effects.

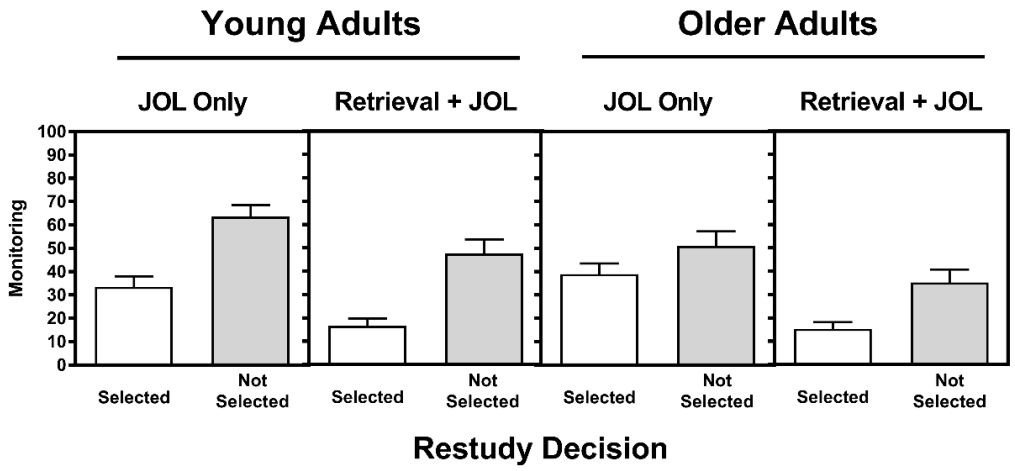

Figure 6. Average judgment of learning rating by item selection, for young and older adults in two monitoring groups in Experiment 3.

In the Figure 6., the interaction between item selection and age group is demonstrated by the difference in average JOL between items selected and not selected to be restudied. Namely, for younger adults, there is a larger difference in average JOL between items selected and not selected to be restudied. For older adults, while a similar pattern is present (items selected for restudy are rated, on average, lower), there is not the same degree of difference, results which parallel those from Experiment 1.

Unlike Experiment 1, there is a marked difference in average JOL between the monitoring groups. Specifically, individuals within the Retrieval + JOL groups, demonstrated lower average monitoring ratings. This pattern was observed for both young and older adults, and for both items selected and not selected to be restudied. This difference in confidence in the future recallability of items can be seen also in the frequency distribution displayed in Figure 7.

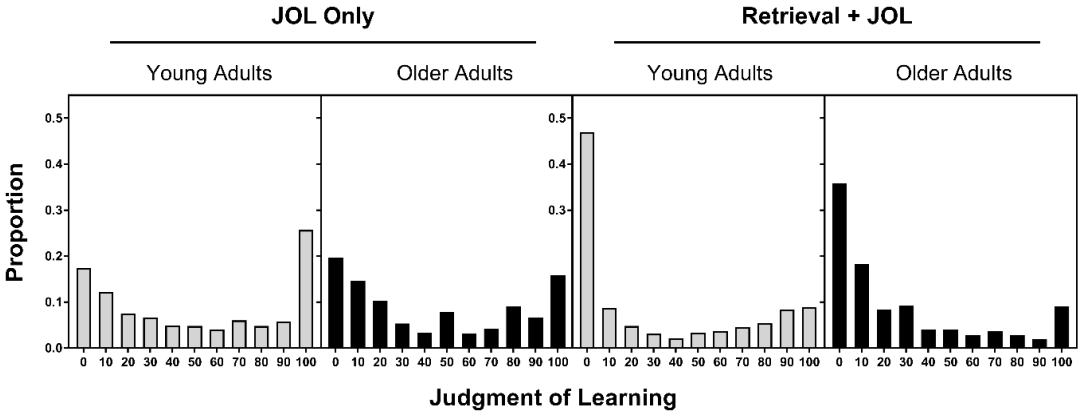

Figure 7. Frequency distribution of the use of each of the JOL categories, plotted separately by groups. Frequency is relative (proportional) as young and older adults received a different number of trials.

In the above figure, one can see the differences in the frequency various monitoring ratings were used. Notably, individuals in the Retrieval + JOL groups used confidence ratings of '0' to a greater extent than individuals in the JOL only groups. The

differences in the frequency of confidence ratings further illustrates how differences in magnitude were present between the monitoring groups.

Intra-individual correlations. Gamma correlations were computed between JOL and item selection decisions. Note that *G* could not be computed for 1 older adult as JOL did not vary. In accordance with findings from Experiment 1 and 2, participants were more likely to select an item for restudy that they had assigned lower as opposed to higher monitoring value. There was a significant age difference, $F(1,104) = 5.73$, $p = .019$. For young adults, average gamma correlations for item selection decisions were strong ($M = -.60$, $SD = .46$); whereas for older adults, average gamma correlations were not as strong ($M = -.35$, $SD = .62$). No other effects were significant, Fs<1.

Final Test

We evaluated group differences in final test performance. Mean test performance was calculated for each individual. Due to a programming error 27 older adults were not tested on one final word pair (window – damage). These trials were not included within this, and the subsequent, analysis. A 2 (age group: young, old) X 2 (monitoring: retrieval + JOL, JOL only) between-subjects ANOVA was performed on mean final test performance. No group differences were found. The effect of age was marginally significant, $F(1,104) = 3.61$, $p = .060$. Both older adult groups scored within the 40 – 50% range ($M_{\text{Retrieval + JOL}} = .44$, $SD = .21$, $M_{\text{JOL only}} = .46$, $SD = .22$). Younger adult groups scored within the 50 – 60% range ($M_{\text{Retrieval + JOL}} = .50$, $SD = .27$, $M_{\text{JOL only}} = .57$, $SD = .21$). Neither the effect of monitoring group ($F(1,104) = 1.40$, $p = .24$), or its interaction with age group ($F < 1$), were statistically significant.

We also evaluated group differences in the relationship between final test performance and JOL. We computed *G* correlations for individuals using JOL and whether the item was recalled at final test, for only those items that had been excluded from the *opportunity* to restudy. Note that *G* could not be computed for 7 younger adults and 13 older adults, either because final test performance was a constant for these items or both final test performance and JOL were a constant for the items. A 2 (age group: young, old) X 2 (monitoring: retrieval + JOL, JOL only) between-subjects ANOVA was performed on mean gamma correlation. No significant effects were found, Fs < 1. Average gamma correlations were strong and positive, indicating that participants' JOLs accurately predicted final test performance (see Table 1 for descriptive statistics).

Discussion

Results from Experiment 3 differed from those of the previous experiments in an intriguing way. Namely, when the relationship between monitoring and control was investigate via magnitude, monitoring group differences emerged; however, when investigated via gamma, only age group differences emerged. That is, the manipulation of monitoring (retrieval was explicitly instructed versus not) affected mean confidence in the future recallability of items, for both young and older adults. As was seen in Figure 7, for the Retrieval + JOL groups, the JOL rating of '0' was the most frequently used, which depressed average confidence ratings for young and older adults alike. Importantly, this lower absolute relationship did not manifest in a weaker *relative* relationship. There, older adults did show a weaker relationship, but, were not differentially affected by the monitoring manipulation.

Next, while the reaction time data only partially replicated findings from Son and Metcalfe (2005), reaction time differences were observed between the participant groups. Older adults, in general, took longer than younger adults during the monitoring phase. Individuals in the JOL only group, in general, took less time during the monitoring phase than individuals in the Retrieval + JOL groups. Our manipulation therefore produced differences in the duration of the retrieval search. We suggest that these observed differences (longer search times) altered the magnitude of confidence ratings, in line with the *control-affects-monitoring* hypothesis (e.g. Koriat et al., 2006). Here, attempted retrieval varies in effort or difficulty. Difficulty in retrieving the answer (i.e. slower retrieval times) affects individuals' confidence in future retrieval. Retrieval serves as a metamnenomic cue, and its diagnosticity may be accurate or not (see Benjamin, Bjork, and Schwartz, 1998). In our study, spending more time attempting to retrieve the answer (the Retrieval + JOL groups) depressed confidence ratings, but did not result in changes in monitoring accuracy or test performance.

Further, the differences in the procedure of Experiment 3 produced findings that supported design decisions implemented in the previous experiments. Specifically, prior to the selection phase, individuals indicated how many items they wanted to restudy. This was done to ascertain how many cues participants would select if able to do so without constraint. We could then compare this number to the number we allowed individuals to select. Additionally, we withheld a third of the cues from the selection phase. To test monitoring prediction accuracy in the absence of the choice to restudy, one-third of the cues were excluded from the restudy phase and prediction accuracy for these items was calculated and compared across groups. First, individuals on average indicated wishing to

restudy half of the items (as was done in the first two experiments). Older adults, as a group, indicated wanting to restudy more items; this pattern seemed to be driven largely by those older adults in the JOL only group. Second, prediction accuracy was high and was equal across groups. Further, monitoring accuracy calculated in this study for items excluded from the restudy stage appeared comparable to monitoring accuracy calculated in the previous experiments (for items studied only once) reported in Table 1. In general, these findings suggest that limiting choices did not unduly influence what individuals' selection behavior would have been and that differences observed between the age groups in the relationship between monitoring and control were not attributable to the accuracy of monitoring.

General Discussion

The present study produced several important findings. First, in Experiment 1, age differences were found in the relationship between monitoring and the control of subsequent study *only* for those items that had not been successfully retrieved at the initial test. Specifically, older adults did not exhibit as strong a relationship as young adults between monitoring and control. Second, in Experiment 2, manipulating the demands of the control task impacted both young and older adults, albeit in different ways. For young adults, reliance on access to partial information increased. For older adults, monitoring and control were no longer related. Finally, in Experiment 3, explicit retrieval influenced monitoring magnitude *only*. Young and older adults gave lower absolute confidence ratings when asked to explicitly retrieve (versus not). The instruction to retrieve did not influence the *relative* relationship between monitoring and control;

however, as in the previous two experiments, older adults did not exhibit as strong a relationship as young adults.

Retrieval as the Basis for Monitoring

Results from Experiment 1 found that control decisions were based largely on one's feeling of knowing. Judgments made at retrieval, such as FOKs, rely on the partial information retrieved (Koriat, 1993), consistent with the results of the present experiments. Importantly, in our studies, the groups were equal in the frequency of retrieving correct partial information. Though initial test performance differed between participant groups, accessing partial information did not. Further, the predictive accuracy of FOK was extremely high for all participant groups (see Table 1). Both younger and older adults were similarly accurate at predicting future memory performance using feeling of knowing and also judgments of learning (in Experiment 3). Based on these findings, we do not believe that age differences found in the relationship between monitoring and control were driven by either recall or monitoring accuracy.

We hypothesize that age differences in these studies may in part be driven by the difficulty of the metamemorial task. Nelson and Leonesio (1988) proposed that monitoring the relative ease of learning or recalling different items is the basis for control. Control, or strategy use, is implemented to compensate for differences in assessed initial learning. Thus, monitoring and control rely heavily on assessed learning and retrieval effort (Koriat et al., 2006). In instances where a priori item difficulty can be used as a cue for monitoring and control, younger and older adults have exhibited identical selection behavior (Dunlosky & Hertzog, 1997; Tullis & Benjamin, 2012). In the second case, participants studied word pairs that were composed of either two abstract

words or two concrete words. Not surprisingly, all participants elected to restudy more abstract—abstract word pairs, the objectively more difficult items. However, when item difficulty is held constant, participants are required to rely on different cues to make monitoring and control decisions. In the current study, individuals could not rely on salient item characteristics (i.e. concreteness or relatedness), as all words were unrelated. In these instances, attempted retrieval and its cues may be important, and may also be evaluated differently by younger and older adults.

In the first two experiments, we required all learners to overtly attempt retrieval immediately prior to engaging in monitoring. Measuring objective performance enabled us to test whether the relationship between monitoring and control differed as a function of retrieval success. Indeed, in Experiment 1, the relationship between monitoring and control differed for young and older adults, but only for those items not recalled at the initial test. This age difference may have been driven by changes in confidence pursuant to the test taking *experience*. For example, experience taking overt tests may lead to reduced confidence in future recallability of the tested items (Koriat, Sheffer, & Ma'ayan, 2002). This *underconfidence with practice* effect has been demonstrated during multi-trial learning, where the experience of taking an overt test affects monitoring that takes place after the test has concluded. Further support for this position was found in Experiment 3. Individuals instructed to explicitly attempt retrieval prior to engaging in monitoring demonstrated lower absolute confidence than those not instructed to attempt retrieval. This manipulation affected young and older adults equally.

For older adults, as it is less likely that they spontaneously engage in practices that could enhance memory performance (e.g. Craik, Klix, & Hagendorf, 1986), one

might consider that explicitly requiring retrieval may be particularly advantageous for older adults compared to younger adults. In particular, self-paced testing is one of the few training methods successful in improving older adult associative learning (Dunlosky, Kubat-Silman, & Hertzog, 2003). However, results from our study suggest that there is not an inherent benefit to the act of explicitly requiring retrieval. Older adults demonstrated weaker relationship relationships between monitoring and control than younger adults when asked to retrieve *and also* when not asked to retrieve. Thus, though requiring learners to explicitly attempt retrieval resulted in lower confidence ratings (Miller & Geraci, 2014), no differences were observed in the number of words individuals wished to restudy (Kimball, Smith, & Muntean, 2012), in the relative relationship between monitoring and control, or in final test performance (Son & Metcalfe, 2005). In this regard, explicit retrieval offered no benefit in the present study. However, as will be discussed shortly, perhaps the benefit would emerge only in certain contexts.

Our findings support the idea that task design affected strategic control. Namely, average FOK predicted the selection of unrecalled and recalled items in Experiment 1 but not in Experiment 2. Specifically, individuals with higher overall confidence in the future recallability of items (higher FOK) were more likely to select unknown items and were less likely to select known items (items correct at initial test). That strategy implementation varied by retrieval success suggests individuals considered goals or how one would aim to complete the task as a whole. That is, individuals with higher average confidence chose items they had not successfully retrieved (the objectively more difficult items); whereas individuals with lower average confidence were more likely to choose

the objectively easier items (items successfully recalled). This differential strategy implementation parallels results examining the relationship between working memory capacity and the regulation of study. For example, working memory capacity (WMC) has been shown to be associated with control behaviors in older adults such that individuals with low WMC tend to select easier items to study and may even avoid difficult items entirely (Price, Hertzog, & Dunlosky, 2010). Individual differences in WMC predicted strategy implementation. In Experiment 1, people who exhibited lower average feeling of knowing were more likely to select items they had successfully retrieved and less likely to select items they had not successfully retrieved, regardless of age. Average FOK was a characteristic of an individual which predicted strategy use.

We did not see evidence of differential strategy implementation in Experiment 2 (where task design depressed the relationship between monitoring and control) or in Experiment 3 (where task design depressed monitoring magnitude). In Experiment 3, people who exhibited lower average confidence (participants in the Retrieval + JOL groups) did not differ in their selection behavior than individuals whose confidence ratings were not artificially depressed. However, all items used within the experiment were difficult (i.e. unrelated). It is possible that manipulating average confidence (via explicit testing) results in different selection behavior *only* when items vary in a priori or objective difficulty. This position requires that objective difficulty be *salient* for artificially lowering confidence to result in different strategy implementation. Future studies could examine this possibility.

The Impact of the Demands of the Task

In Experiment 1, we examined age differences in metacognitive control in task conditions that promoted goal-driven processes in strategy selection (Dunlosky & Hertzog, 1997). Specifically, participants chose items to restudy from an array that presented all studied cues; thus, item selection occurred during one phase, which was separate from the initial testing and monitoring phase. In Experiment 2, participants made selection decisions during the initial testing phase, and, therefore, had to concurrently manage several ongoing processes. The effects of this change were pronounced. As measured by intra-individual correlations, the relationship between FOK and restudy was not different than zero for older adult participants. Though young adults did demonstrate an average intra-individual correlation between FOK and selection that was different than zero, results from the mixed effects model indicated a more significant predictor of restudy: partial information. Successfully retrieved partial information of an unrecalled target was associated with a *decrease* in likelihood of selecting an item to be restudied, for young adults only. Interestingly, this only appeared in Experiment 2, where the use of goal-driven processes during the control task were reduced, and the likelihood of relying on data-driven processes increased. For young adults, this suggests a self-regulated learning pattern consistent with the discrepancy reduction model. However, partial information could cue that an item is *closer to being learned*, and therefore may signal that the item is in one's region of proximal learning. A possible future direction might include instructing individuals to make decisions in line with that cue (accessing partial information), and testing the effectiveness of those decisions (by for example, honoring or dishonoring people's choices, c.f. Kornell & Metcalfe, 2006).

Finally, it is important to note that the participants in this experiment represented high-functioning older adults with good verbal knowledge. Indeed, as a group, the older adults had better vocabulary scores and were more highly educated than their younger adult counterparts. With less cognitively able older adults one might expect even greater effects of age on metacognitive control. One other limitation should be considered. It is important to acknowledge that *two* method changes were made between Experiments 1 and 2, that of presentation *and* timing. Because of this, it cannot be known to what extent monitoring and control was affected by the serial, one at-a-time presentation of the choice or by the choice being made during the testing phase. A control condition would therefore be the serial, one at-a-time presentation of choices *after* testing has concluded. In that instance, some time will have elapsed between attempted recall and monitoring, which may result in a different pattern of metacognitive control. For example, because the initial attempt to retrieve and control would no longer be in such direct proximity, retrieval success may no longer be as strong a predictor of control (as was demonstrated in Experiment 2).

Conclusions

This study offered several unique features. First, we considered self-regulated learning within the context of retrieval. Because retrieval was explicitly measured, we were able to specify where age differences emerged. In Experiments 1 and 2, differences in the relationship between monitoring and control were found for unrecalled items only. Importantly, young and older adults demonstrated similarities. When items were presented in an array, young and older adults adopted similar strategies. When retrieval was explicitly measured and successful, both young and older adults used FOK to guide

subsequent control. However, across experiments older adults demonstrated a weaker relationship between monitoring and control. This finding persisted whether attempted retrieval was required or not. This signals that cues associated with retrieval monitoring may be used differently by young and older adults, and may be more nuanced than previously thought.

Table 1

Average within-subject gamma correlations (with *SD* in parentheses) between FOK and 1) restudy choice and 2) final test (for items studied only once in Experiments 1 and 2).

		Experiment 1		Experiment 2	
		Young Adults	Older Adults	Young Adults	Older Adults
Restudy	Valence First	- .58 (*.31*)	- .29 (*.59*)	- .48 (*.45*)	- .17 (*.60*)
	FOK First	- .53 (*.44*)	- .18 (*.58*)		
Final Test	Valence First	.82 (*.20*)	.85 (*.24*)	.89 (*.19*)	.72 (*.49*)
	FOK First	.80 (*.35*)	.62 (*.65*)		
		Experiment 3			
Final Test	JOL Only	.79 (*.44*)	.82 (*.32*)		
	Retrieval + JOL	.92 (*.18*)	.81 (*.30*)		

Table 2

Summary of mixed-effects binomial logistic regression predicting likelihood of restudy

	Items not correctly recalled		Items correctly recalled	
Experiment 1				
Fixed Effects	β (SE)	Exp(β)	β (SE)	Exp(β)
Intercept	0.37 (.08)***	1.45	- 0.77 (.27)**	0.46
Age (older)	- 0.05 (.10)	0.95	0.65 (.38)$^{\sim}$	1.91
$FOK_{between}$[1]	0.15 (.04)***	1.16	- 0.28 (.12)*	0.76
$FOK_{between}$ × Age (older)	- 0.01 (.05)	0.99	- 0.13 (.16)	0.88
FOK_{within}[2]	- 0.24 (.03)***	0.79	- 0.35 (.07)***	0.70
FOK_{within} × Age (older)	0.16 (.05)***	1.17	0.04 (.10)	1.04
Random Effects				
Item random intercept variance	0.00 (.00)		0.04 (.04)	
Subject random intercept variance	0.00 (.00)		1.40 (.11)	
Experiment 2				
Fixed Effects	β (SE)	Exp(β)	β (SE)	Exp(β)
Intercept	1.84 (.23)***	6.19	- 1.04 (.50)*	0.35
Age (older)	- 0.64 (.31)*	0.54	1.09 (.72)	2.97
FOK_{within}	- 0.03 (.06)	0.96	- 0.95 (.20)***	0.39
FOK_{within} × Age (older)	- 0.03 (.10)	1.18	- 0.39 (.32)	0.68
Partial Information	- 1.19 (.25)***	0.31		
Partial Information × Age	1.34 (.35)***	3.69		
Random Effects				
Item random intercept variance	0.01 (.02)		0.07 (.05)	
Subject random intercept variance	0.56 (.09)		2.76 (.15)	

Note: $^{\sim}p$ <.10 *p < .05. **p < .01. ***p < .001.

Table 3

Goodness-of-fit statistics and χ^2 for models predicting reaction time in milliseconds.

JOL Only	– 2LL	BIC	Δ *df*	Δ χ^2
1. Baseline	24894	24923		
2. JOL_{trial}	24892	24928	1	2.16
3. JOL_{trial}^2	24793	24836	1	98.97***
4. Age and Age interactions	24750	24816	3	43.00***
Retrieval + JOL	– 2LL	BIC	Δ *df*	Δ χ^2
Retrieval				
1. Baseline	25529	25557		
2. JOL_{trial}	25418	25454	1	110.43***
3. JOL_{trial}^2	25376	25429	1	42.06***
4. Age and Age interactions	25427	25362	3	13.80**
JOL				
1. Baseline	22477	22444		
2. JOL_{trial}	22408	22418	1	68.62***
3. JOL_{trial}^2	22295	22338	1	113.66***
4. Age and Age interactions	22286	22351	3	08.96*

Note: *$p < .05$. **$p < .01$. ***$p < .001$.

Table 4

Summary of mixed-effects generalized linear model predicting reaction time

JOL Only	**JOL**
Fixed Effects	β (SE)
Intercept	3250.88 (278.2)***
Age (older)	1510.60 (416.5)***
JOL_{trial}	322.70 (68.3)***
JOL_{trial} × Age (older)	732.24 (148.5)***
JOL_{trial}^2	- 33.43 (6.6)***
JOL_{trial}^2 × Age (older)	- 74.27 (104.5)***
Random Effects	
Item random intercept variance	0.00 (28.7)
Subject random intercept variance	0.00 (181.2)

Retrieval + JOL	**Retrieval**	**JOL**
Fixed Effects	β (SE)	β (SE)
Intercept	5236.08 (291.2)***	1257.96 (123.4)***
Age (older)	2436.49 (451.0)***	695.33 (182.6)***
JOL_{trial}	399.22 (106.8)***	375.97 (43.3)***
JOL_{trial} × Age (older)	- 393.61 (200.4)*	- 29.61 (79.4)
JOL_{trial}^2	- 62.15 (11.1)***	- 35.46 (4.6)***
JOL_{trial}^2 × Age (older)	36.29 (20.3)$^{\sim}$	4.59 (8.4)
Random Effects		
Item random intercept variance	0.00 (182.6)	0.00 (17.4)
Subject random intercept variance	0.00 (25.2)	0.00 (81.6)

Note: $^{\sim}p$ <.10 *p < .05. **p < .01. ***p < .001.

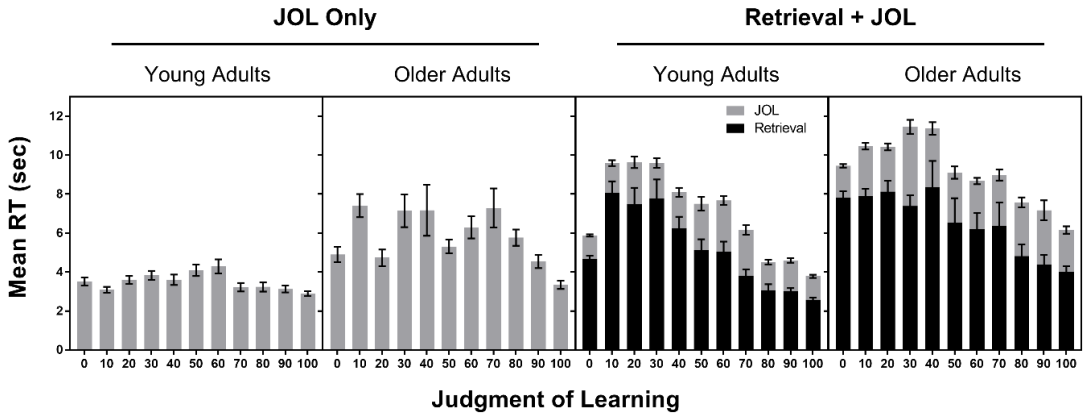

Figure 5. Mean reaction times (RT) in each of the JOL categories. In the Retrieval + JOL condition, reaction times to the first button-press (retrieval) are in black. For both conditions, reaction times to the JOL screen are in gray.

Bibliography

American Psychological Association. (2014). Guidelines for Psychological Practice with Older Adults. *American Psychologist*, *69*(1), 34–65. https://doi.org/10.1037/a0035063

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68. https://doi.org/http://dx.doi.org/10.1037/0096-3445.127.1.55

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73–94). New York: Psychology Press.

Brewer, G. A., Marsh, R. L., Clark-Foos, A., & Meeks, J. T. (2010). Noncriterial recollection influences metacognitive monitoring and control processes. *The Quarterly Journal of Experimental Psychology*, *63*(10), 1936–1942. https://doi.org/10.1080/17470210903551638

Castel, A. D., Benjamin, A. S., Craik, F. I., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory & Cognition*, *30*(7), 1078–1085.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates. Hillsdale, NJ, 20–26.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Cook, G. I., Marsh, R. L., & Hicks, J. L. (2006). Source memory in the absence of successful cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 828–835. https://doi.org/http://dx.doi.org/10.1037/0278-7393.32.4.828

Craik, F. I. M. (2000). Age-related changes in human memory. In D. C. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 75–92). New York, NY: Psychology Press.

Craik, F. I., Klix, K., & Hagendorf, H. (1986). A functional account of age differences in memory. In *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409–422). Amsterdam: Elsevier, North-Holland.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition*, *25*(5), 691–700.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52*(4), P178–P186.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 249–276). Mahwah, NJ: Erlbaum.

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Training monitoring skills improves older adults' self-paced associative learning. *Psychology and Aging*, *18*(2), 340–345. https://doi.org/10.1037/0882-7974.18.2.340

Dunlosky, J., Serra, M. J., & Baker, J. M. (2007). Metamemory applied. In F. T. Durso (Ed.), *Handbook of applied cognition* (Vol. 2, pp. 137–161). Chichester; New York: John Wiley & Sons.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica*, *98*(1), 37–56. https://doi.org/10.1016/S0001-6918(97)00051-6

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Grober, E., Sliwinsk, M., & Korey, S. R. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, *13*(6), 933–949. https://doi.org/10.1080/01688639108405109

Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and "don't know" responding in episodic memory tasks. *Journal of Memory and Language*, *69*(3), 368–383. https://doi.org/10.1016/j.jml.2013.04.005

Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic Bulletin & Review*, *21*(6), 1617–1622. https://doi.org/10.3758/s13423-014-0619-0

Hertzog, C. (2016). Aging and metacognitive control. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 537–558). Oxford University Press.

Hertzog, C., Dunlosky, J., & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition*, *38*(6), 771–784. https://doi.org/10.3758/MC.38.6.771

Hertzog, C., Fulton, E. K., Sinclair, S. M., & Dunlosky, J. (2014). Recalled aspects of original encoding strategies influence episodic feelings of knowing. *Memory & Cognition, 42*(1), 126–140. https://doi.org/10.3758/s13421-013-0348-z

Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: an analysis of adult age differences. *Psychology and Aging, 24*(2), 462–475. https://doi.org/10.1037/a0014417

Jacoby, L. L., & Rhodes, M. G. (2006). False remembering in the aged. *Current Directions in Psychological Science, 15*(2), 49–53.

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 923–954. https://doi.org/http://dx.doi.org/10.1037/a0026936

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36–69. https://doi.org/10.1037/0096-3445.135.1.36

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147–162. https://doi.org/10.1037//0096-3445.131.2.147

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 609–622. https://doi.org/http://dx.doi.org/10.1037/0278-7393.32.3.609

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Littrell, M. K. (2011). *The influence of testing on memory, monitoring, and control* (Ph.D.). Colorado State University, United States -- Colorado. Retrieved from https://search.proquest.com/docview/889127066/abstract/635212432C344529PQ/1

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

Marsh, E. J., Dolan, P. O., Balota, D. A., & Roediger, H. L. (2004). Part-set cuing effects in younger and older adults. *Psychology and Aging*, *19*(1), 134–144. https://doi.org/http://dx.doi.org/10.1037/0882-7974.19.1.134

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*(3), 349–363. https://doi.org/http://dx.doi.org/10.1037/0096-3445.131.3.349

Miles, J. R., & Stine-Morrow, E. A. L. (2004). Adult age differences in self-regulated learning from reading sentences. *Psychology and Aging*, *19*(4), 626–636. https://doi.org/http://dx.doi.org/10.1037/0882-7974.19.4.626

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*(Supplement C), 131–140. https://doi.org/10.1016/j.concog.2014.08.008

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/http://dx.doi.org/10.1037/a0036914

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved February 9, 2017, from http://w3.usf.edu/FreeAssociation/

Nelson, H. E., & Willison, J. (1991). *National Adult Reading Test (NART).* Nfer-Nelson Windsor. Retrieved from http://www.academia.edu/download/31611053/NART_MANUAL.pdf

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–270.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676–686. https://doi.org/http://dx.doi.org/10.1037/0278-7393.14.4.676

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bowers (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment Recall and Monitoring (PRAM). *Psychological Methods*, *9*(1), 53–69. https://doi.org/10.1037/1082-989X.9.1.53

Price, J., Hertzog, C., & Dunlosky, J. (2010). Self-regulated learning in younger and older adults: Does aging affect metacognitive control? *Aging, Neuropsychology, and Cognition*, *17*(3), 329–359. https://doi.org/10.1080/13825580903287941

Price, J., & Murray, R. G. (2012). The region of proximal learning heuristic and adult age differences in self-regulated learning. *Psychology and Aging*, *27*(4), 1120–1129.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Salthouse, T. (2010). *Major issues in cognitive aging*. Oxford University Press.

Schwartz, B. L., Boduroglu, A., & Tekcan, A. İ. (2016). Methodological concerns: the feeling-of-knowing task affects resolution. *Metacognition and Learning*, *11*(3), 305–316. https://doi.org/10.1007/s11409-015-9152-4

Schwartz, B. L., & Metcalfe, J. (1994). Methodological Probems and Pitfalls in the Study of Human Metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). MIT Press.

Schwartz, B. L., Pillot, M., & Bacon, E. (2014). Contextual information influences the feeling of knowing in episodic memory. *Consciousness and Cognition*, *29*, 96–104. https://doi.org/10.1016/j.concog.2014.08.018

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204–221.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*, *33*(6), 1116–1129.

Souchay, C., & Isingrini, M. (2004). Age-related differences in the relation between monitoring and control of learning. *Experimental Aging Research*, *30*(2), 179–193. https://doi.org/10.1080/03610730490274248

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1024–1037. https://doi.org/http://dx.doi.org/10.1037/0278-7393.25.4.1024

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2011). Context affects feeling-of-knowing accuracy in younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 96–108. https://doi.org/10.1037/a0021612

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2012). An analysis of the determinants of the feeling of knowing. *Consciousness and Cognition*, *21*(4), 1681–1694. https://doi.org/10.1016/j.concog.2012.09.005

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review*, *19*(4), 743–749. https://doi.org/10.3758/s13423-012-0266-2

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Zachary, R. A. (1991). *Shipley institute of living scale*. WPS, Western Psychological Services.